On the Minimal Degree Bias in Generalization on the Unseen for non-Boolean Functions

Denys Pushkin [1] **Raphaël Berthier** [2] **Emmanuel Abbe** [1 3]

Abstract

We investigate the out-of-domain generalization of random feature (RF) models and Transformers. We first prove that in the 'generalization on the unseen (GOTU)' setting, where training data is fully seen in some part of the domain but testing is made on another part, and for RF models in the small feature regime, the convergence takes place to interpolators of minimal degree as in the Boolean case (Abbe et al., 2023). We then consider the sparse target regime and explain how this regime relates to the small feature regime, but with a different regularization term that can alter the picture in the non-Boolean case. We show two different outcomes for the sparse regime with q-ary data tokens: (1) if the data is embedded with roots of unities, then a min-degree interpolator is learned like in the Boolean case for RF models, (2) if the data is not embedded as such, e.g., simply as integers, then RF models and Transformers may not learn minimal degree interpolators. This shows that the Boolean setting and its roots of unities generalization are special cases where the minimal degree interpolator offers a rare characterization of how learning takes place. For more general integer and real-valued settings, a more nuanced picture remains to be fully characterized.

1. Introduction

Some of the most challenging tasks for state-of-the-art machine learning models reside in settings where the training data is not representative of the testing data, or more specifically, when there is significant distribution shift. This is in particular central to 'reasoning tasks', such as arithmetic and algebra (Saxton et al., 2019; Lewkowycz et al., 2022), visual reasoning such as CLEVR (Johnson et al., 2017), physical reasoning such as Phyre (Bakhtin et al., 2019), algorithmic data such as CLRS (Veličković et al., 2022) and reasoning on graphs (Mahdavi et al., 2022). In such settings, the combinatorial nature of the data makes comprehensive data sampling challenging, resulting naturally in 'holes' in the sampled domain.

An archetype example of this kind is the 'length generalization' setting: no matter how dense we sample discrete inputs, the training data will have inputs of some bounded length, and one can naturally ask for generalization to larger length. This has motivated Abbe et al. (2023) to consider a special case of out-of-distribution generalization: generalization on the unseen (GOTU). In its most extremal form, the GOTU setting assumes that part of the domain (in a large embedded dimension) is fully observed at training, and the generalization of the model is tested on a new (unseen) part of the domain. Therefore in this setting, there is no 'estimation error' since the model always learns perfectly in-distribution, but the question of interest is to understand how well the model generalizes on new domains depending on the model parametrization and the optimization.

In (Abbe et al., 2023), it is shown that for sparse Boolean functions, i.e., functions defined on $\{+1, -1\}^d$ that depend only on a bounded number of variables, and in such a GOTU setting where training data is available in $\mathcal{U}^c \subseteq \{+1, -1\}^d$ and generalization is tested on $\mathcal{U}$, random feature models learn functions that are interpolators on $\mathcal{U}^c$ with minimal degree-profile: a specific type of polynomials that have minimal degree and also largest mass on the lowest-degree Fourier coefficients. Further, experiments provided in (Abbe et al., 2023) show that this 'minimal degree bias' also takes place for Transformers.

In this paper, we study how this picture changes when con-

[1]School of Computer and Communication Sciences, EPFL, Lausanne, Switzerland [2]Inria Sorbonne Université, Paris, France. (While RB is currently affiliated with Inria, the work presented here was partly done while affiliated with the School of Computer and Communication Sciences, EPFL, Lausanne, Switzerland) [3]Apple, Machine Learning Research (MLR), Switzerland. Correspondence to: Denys Pushkin <denys.pushkin@epfl.ch>, Raphael Berthier <raphael.berthier@inria.fr>, Emmanuel Abbe <emmanuel.abbe@epfl.ch>.

Proceedings of the 41st International Conference on Machine Learning, Vienna, Austria. PMLR 235, 2024.

sidering input variables that are not necessarily valued in $\{+1,-1\}^d$. In particular, we consider input variables valued in $\mathcal{X} \subseteq \mathbb{R}^d$, which may be discrete but non-binary or arbitrary real-valued. We study theoretical results for random feature models and provide experiments for Transformers.

We make this study in two settings: (i) the sparse regime as in (Abbe et al., 2023), where the target function depends effectively on a low input dimension, (ii) the small feature regime, where random features have a weight scale that tends to zero while the input dimension remains fixed. In particular, we explain how these two regimes are related to each other, with (ii) acting as a surrogate of (i) with a regularization term, and investigate both regimes. We further provide experiments for Transformers.

2. Paper Contributions and a Motivating Example

Consider the arithmetic task where inputs are valued in $\mathcal{X} = \{-q/2, \dots, -1, 1, \dots, q/2\}$, for some fixed even q, and the target function $f : \mathcal{X}^d \to \mathcal{X} \cdot \mathcal{X}$ is given by

$$f(x_1, x_2, \dots, x_d) = x_1 \cdot x_2 \,.$$

In the GOTU setting, we assume that a set of training examples is given that covers some part of the support and leaves out completely some other part. For instance, consider the case where x_1 or x_2 is always 1 at training. This is a special case where the model would a priori not have a reason to learn the target function (it does not see effective multiplications). So what will the model learn in that case?

Note that one can model the GOTU constraint in this case as follows:

$$(x_1 - 1)(x_2 - 1) = 0 \quad \equiv \quad x_1 x_2 = x_1 + x_2 - 1 \,.$$

One possible outcome is that the model could learn a function close to $\widehat{f}(x) = x_1 + x_2 - 1$. This is explained by the following intuition: this function is a correct interpolator of the training data, and it has lowest possible degree. Assuming that such models have a bias towards lower degree polynomials, this function may be learned. This turns out to be a correct intuition in the Boolean case, i.e., when $q = 2$. More precisely, this was proved by Abbe et al. (2023) for classic RF models when d diverges, i.e., the 'sparse regime', and experiments supporting a similar outcome for Transformers were also obtained. In this paper, as a first contribution, we show that this outcome also takes place when the target is not sparse (e.g., $d = 2$) but the random features have a vanishing variance, which we call the 'small feature regime'; this in fact provides a surrogate regime to the sparse regime.

What happens now if $q > 2$? Would such models still learn a function close to $\widehat{f}(x) = x_1 + x_2 - 1$? In this paper, we show the following:

1. (Small-feature RF and any target) Yes, this intuition is still correct for RF models in the small feature regime and any real inputs. See Theorem 4.2 for a formal statement.

2. (Sparse target on arbitrary inputs) No, this intuition is incorrect for classic RF models with general activations on sparse targets, as the model can learned higher degree polynomials, although some activations such as sigmoid appear to still obey the minimal degree rule; similarly, this intuition is not correct in general for Transformers, as we provide experiments with both minimal degree and higher degree interpolators. These experiments are reported in Section 7.

3. (RF and sparse target on roots of unities) Yes, this intuition is again correct for classic RF models with general activations (with some regularity condition) and sparse targets if the data is not parametrized as $\mathcal{X} = \{-q/2, \dots, -1, 1, \dots, q/2\}$ but as $\mathcal{U} = \{e^{2\pi ik/n}\}_{k=0}^{n-1}$, with the same target $x_1 \cdot x_2$ over the complex numbers (i.e., the target is now the sum of angles of the roots of unities). This is the 'natural' extension of the Boolean case (with $q = 2$) to larger q. Note that this is not due to the fact that the target here becomes the sum of the angles, as the result extends to more generic functions. We leave it to future work to investigate whether this parametrization could be useful in certain applications; it may also generalize to other groups than roots of unities. We refer to Theorem 6.1 in Section 6 for the formal statement.

3. Background

3.1. Notation

We denote $\mathbb{N}$ the set of non-negative integers. If $T \in \mathbb{N}^d$, we denote $|T| = \|T\|_1 = T_1 + \dots + T_d$. We define $\mathbb{N}^d_{\leqslant p} = \{T \in \mathbb{N}^d \,|\, |T| \leqslant p\}$. We also denote $\mathbb{U}_n = \left\{\exp\left(i\frac{2\pi k}{n}\right) \mid k = 0, \dots, n-1\right\} \subset \mathbb{C}$ the n-roots of unity.

We denote $\Pi_p(\mathbb{R}^d)$ the set of multivariate real polynomials on $\mathbb{R}^d$ with degree less or equal to p. Similarly, we denote $\Pi_p(\mathbb{U}_n^d)$ the set of functions on $\mathbb{U}_n^d$ that are the restriction to $\mathbb{U}_n^d$ of a multivariate complex polynomial on $\mathbb{C}^d$ of degree at most p.

If $(V, \|.\|)$ is a normed vector space, $x \in V$ and W is a subspace of V, then $\mathrm{dist}(x, W)$ denotes the distance of x to W. We denote γ_d the standard Gaussian measure over $\mathbb{R}^d$.

3.2. Random Feature Model and Different Regimes

In the following sections, we will study the random features (RF) model $f_{\mathrm{RF}}(a) : \mathbb{R}^d \to \mathbb{R}$, which is defined as $f_{\mathrm{RF}}(a; x) = \frac{1}{\sqrt{N}} \sum_{i=1}^N a_i \phi_{w_i, b_i}(x)$, $x \in \mathbb{R}^d$.

Here, $x \in \mathbb{R}^d$ is the input variable, $a \in \mathbb{R}^N$ are the trainable parameters, and $\phi_{w_i,b_i}(x), i \in \{1, \dots, N\}$ are the random features, defined by $\phi_{w,b}(x) = \sigma(\langle w, x\rangle + b)$, where the weights w_i and biases b_i are sampled randomly and then fixed during the training.

Traditionally, the weights w_i and biases b_i of random features model are sampled independently and identically distributed (i.i.d.) according to $w_i \sim \mathcal{N}(0, \frac{1}{d}I_d)$, $b \sim \mathcal{N}(0, \frac{1}{d})$. (Abbe et al., 2023) analysed this setting with additional assumptions that the target function f must be $O_d(1)$-sparse (i.e. it must depend on the finite number of variables) while the dimension d diverges. We call this setting *the sparse regime*.

Additionally, we consider the setting where $w_i \sim \mathcal{N}(0, \varepsilon I_d)$, $b \sim \mathcal{N}(0, \varepsilon)$. Here, we assume that the dimension d is fixed, but $\varepsilon \to 0$. We call this setting *the small features regime*.

As we will see, the sparse and small features settings are related to each other: we can show that they are equivalent up to some regularizer term (see Section 5). However, we will see that, at least for polynomial activation functions, they have different generalization properties in a GOTU setting. In small features regime, the random features model converges to a minimum-degree interpolator (MDI) (under some general assumption on polynomial activation function, see Section 4), while in the sparse regime the convergence to MDI is rather an exception (see Example 5.3 and Remark 5). Finally, we note that the sparse regime requires dimension d to be large and target function to be sparse. On the other hand, the small features regime does not impose any constraint on dimension d or target function f, but requires non-classical initialization of the weights and biases. Thus, these two setting have different limitations and areas of applicability.

Define the image of f_{RF} model as the set of functions it can express: $\mathrm{im}(f_{\mathrm{RF}}) = \{f_{\mathrm{RF}}(a),\, a \in \mathbb{R}^N\}$.

4. Min-Degree Interpolation in Small Features Regime

Let $\mathcal{U} \subset \mathbb{R}^d$ be the unseen domain, so that during the training we only see samples from $\mathcal{U}^c = \mathbb{R}^d \setminus \mathcal{U}$. We emphasize that being a proper subset of $\mathbb{R}^d$ is the only constraint we impose on the unseen domain $\mathcal{U}$. In particular, we can select $\mathcal{U}$ such that the training domain $\mathcal{U}^c$ is finite or countable with a discrete measure defined on it. Similar to (Abbe et al., 2023), we assume that the model has an access to the distribution on the training domain, which makes sampling error zero and allows to state more clear results.

For the activation function σ we assume the following.

Assumption 4.1. Assume that σ is a polynomial of degree p whose coefficients are non-zero:

$$\sigma(y) = b_p y^p + \ldots + b_1 y + b_0, \quad \text{where } b_p, \ldots, b_0 \neq 0\,.$$

Theorem 4.2. *Consider training the random features model $f_{\mathrm{RF}}(a; x)$ in the small features regime (with parameter ε) on the polynomial target function f. Assume that we observe the target function on the training set $\mathcal{U}^c$, and that the activation function σ satisfies Assumption 4.1.*

For a sufficiently large number N of random features, the model f_{RF} can interpolate the target function perfectly on U^c.

Among all parameters a such that $f_{\mathrm{RF}}(a)$ interpolates f on $\mathcal{U}^c$, denote a^* the parameter of minimum ℓ_2 norm. Denote by p_* the minimum possible degree for a polynomial interpolator of f on the training set $\mathcal{U}^c$. Then with high probability, we have[1]:

$$\lim_{\varepsilon \to 0} \lim_{N \to \infty} \mathrm{dist}(f_{\mathrm{RF}}(a^*), \Pi_{p_*}) = 0\,.$$

Remark 4.3. Note that the model will converge to the minimum ℓ^2 norm solution a^* if trained with (stochastic) gradient descent starting from $a = 0$ initialization under the mean squared error loss (in an overparametrized setting). Thus the theorem describes the bias of gradient descent methods.

Remark 4.4. If the target function f is not a polynomial, we can still apply Theorem 4.2 to describe where the random features model converges. Let λ be the distribution on the training set $\mathcal{U}^c$ and assume that the mean square error is used to train the model. Denote by $\tilde{f}$ the projection of f in $L^2(\mathcal{U}^c, \lambda)$ on the space $\Pi_p(\mathbb{R}^d)$ of polynomials of degree at most p:

$$\tilde{f} = \mathrm{proj}_{\Pi_p(\mathbb{R}^d)}(f) = \mathrm{argmin}_{h \in \Pi_p(\mathbb{R}^d)} \|f - h\|_{L^2(\mathcal{U}^c, \lambda)}$$

Then the loss function can be decomposed as

$$\begin{aligned}
\mathcal{L}(a) &= \mathbb{E}_{x\sim\lambda}\left[(f_{\mathrm{RF}}(a; x) - f(x))^2\right] \\
&= \|f_{\mathrm{RF}}(a) - f\|^2_{L^2(\mathcal{U}^c,\lambda)} \\
&= \|f_{\mathrm{RF}}(a) - \tilde{f} + \tilde{f} - f\|^2_{L^2(\mathcal{U}^c,\lambda)} \\
&= \|f_{\mathrm{RF}}(a) - \tilde{f}\|^2_{L^2(\mathcal{U}^c,\lambda)} + \|\tilde{f} - f\|^2_{L^2(\mathcal{U}^c,\lambda)}\,.
\end{aligned}$$

The second term in the last expression is independent of a. Thus, the training trajectory would be the same as if we trained the model on the target function $\tilde{f} \in \Pi_p(\mathbb{R}^d)$, and we may predict where the random features model converges by applying Theorem 4.2 to the target function $\tilde{f}$.

Remark 4.5. Our theorem is not specific to Gaussian parameters of the random features. The proof also works for any weights and biases of the form $w_i = \varepsilon^{1/2}\bar{w}_i$ and $b_i = \varepsilon^{1/2}\bar{b}_i$, with $\bar{w}_i \sim \mu$ and $\bar{b}_i \sim \nu$ and μ, ν are any distributions with all moments finite.

[1] Since the space $\Pi_d(\mathbb{R}^d)$ has finite dimension, the convergence in all norms is equivalent in this space.

We refer to Appendix B for the proof of the theorem. The proof can be decomposed into two parts. First, we show that for fixed ε, as $N \to \infty$ the random features model, denoted as g, converges to the minimizer of the quadratic form $\hat{g}^T \Phi^{-1} \hat{g}$, denoted as g_ε, where Φ is a feature kernel matrix, and $\hat{g}$ is the vector of Hermite coefficients of g. The proof of this part follows the scheme of Theorem 3.8 from (Abbe et al., 2023). Second, we analyze how this minimizer g_ε behaves in the limit of $\varepsilon \to 0$ and prove that $\mathrm{dist}(g_\varepsilon, \Pi_{p_*}) \to 0$. This part of our proof is original. In the Boolean case of (Abbe et al., 2023), the matrix Φ was diagonal. Hence it was enough to estimate its diagonal entries, which directly leads to the approximation of its inverse. However, in general case matrix Φ is non-diagonal. Thus, we estimate all entries of the matrix Φ and derive the suitable upper and lower bounds on the quadratic form $\hat{g}^T \Phi^{-1} \hat{g}$ (Lemma B.6 and Corollary B.7 in Appendix B). This is the part where most of the technical difficulty and conceptual novelty lies. This is also where the big picture changes with the min-degree bias of (Abbe et al., 2023) breaking, if we do not use the small features regime (see Example 5.3 for the demonstration of min-degree bias breaking).

5. Motivations for the Small Features Setting

In this section, we derive equivalences between the setting with fixed dimension and small features, and the setting with diverging dimension and $O(1)$ features.

Let k denote a fixed dimension and $d \gg 1$ denote a large dimension. We set ourselves in the multi-index model where we seek to learn a function $f : \mathbb{R}^d \to \mathbb{R}$ of the form

$$f(x) = \varphi(U^\top x)\,,$$

where $U \in \mathbb{R}^{d\times k}$, $U^\top U = I_k$ and $\varphi : \mathbb{R}^k \to \mathbb{R}$.

We consider the approximation of f with random features:

$$f_{\mathrm{RF}}(a;x) = \sum_{i=1}^N a_i \sigma(\langle w_i, x\rangle + c_i)\,,$$

where $w_i \sim \mathcal{N}(0, \frac{1}{d} I_d)$ and $c_i \in \mathbb{R}$. (The reasoning actually works for other random features, this is simply to set an order of magnitude for the w_i.)

We define the loss function in the approximation

$$\mathcal{L}(a) = \frac{1}{2}\mathbb{E}_x\left[\left(f(x) - f_{\mathrm{RF}}(a;x)\right)^2\right]\,.$$

Denote $P_\|$ the orthogonal projection onto $\mathrm{im}(U)$ and $P_\perp$ the orthogonal projection onto $\ker U^\top = (\ker U)^\perp$. If $q \in \mathbb{R}^d$, we denote $q_\| = P_\| q$ and $q_\perp = P_\perp q$. We make the following assumption on the input distribution of x.

Assumption 5.1. We assume that $x_\|$ and $x_\perp$ are independent.

Example 5.2. This assumption holds in many cases of interest. We show two examples.

1. If $x \sim \mathcal{N}(0, I_d)$, then the assumptions holds. Indeed, then $x_\| \sim \mathcal{N}(0, P_\|)$ and $x_\perp \sim \mathcal{N}(0, P_\perp)$ are independent.

2. If $x \sim \mathrm{Unif}(\{-1,1\}^d)$ and the columns of U are a subset of the canonical basis. (This means that the multi-index model is sparse, meaning that it only depends on a subset of the coordinates.) In this case, $x_\|$ and $x_\perp$ are independent with uniform distribution on hypercubes of respective dimension k and $d-k$.

Under Assumption 5.1, we compute

$$\begin{aligned}\mathcal{L}(a) &= \frac{1}{2}\mathbb{E}_x\left[\left(f(x) - f_{\mathrm{RF}}(a;x)\right)^2\right] &(1)\\ &= \frac{1}{2}\mathbb{E}_{x_\|}\mathbb{E}_{x_\perp}\left[\left(f(x) - f_{\mathrm{RF}}(a;x)\right)^2\right]. &(2)\end{aligned}$$

As $U^\top x$ is independent of $x_\perp$, we have

$$\begin{aligned}&\mathbb{E}_{x_\perp}\left[\left(f(x) - f_{\mathrm{RF}}(a;x)\right)^2\right]\\ &= \mathbb{E}_{x_\perp}\left[\left(\varphi(U^\top x) - f_{\mathrm{RF}}(a;x)\right)^2\right]\\ &= \left(\varphi(U^\top x) - \mathbb{E}_{x_\perp} f_{\mathrm{RF}}(a;x)\right)^2 + \mathrm{var}_{x_\perp} f_{\mathrm{RF}}(a;x) =\\ &\left(\varphi(U^\top x) - \sum_{i=1}^N a_i \mathbb{E}_{x_\perp}\sigma(\langle w_{i\|}, x_\|\rangle + \langle w_{i\perp}, x_\perp\rangle + c_i)\right)^2\\ &+ \mathrm{var}_{x_\perp}\left(\sum_{i=1}^N a_i\sigma(\langle w_i, x\rangle + c_i)\right).\end{aligned}$$

We denote $\overline{\sigma}_i(\lambda) = \mathbb{E}_{x_\perp}\left[\sigma(\lambda + \langle w_{i\perp}, x_\perp\rangle)\right]$. This corresponds to a smoothed version of the non-linearity σ. (For instance, in the case of Gaussian inputs $x \sim \mathcal{N}(0, I_d)$, the smoothing noise $\langle w_{i\perp}, x_\perp\rangle$ would be Gaussian $\mathcal{N}(0, \|w_{i\perp}\|^2)$). We then obtain:

$$\begin{aligned}&\mathbb{E}_{x_\perp}\left[\left(f(x) - f_{\mathrm{RF}}(a;x)\right)^2\right]\\ &\quad= \left(\varphi(U^\top x) - \sum_{i=1}^N a_i\overline{\sigma}_i(\langle w_{i\|}, x_\|\rangle + c_i)\right)^2\\ &\quad+ \sum_{i,j=1}^N a_i a_j \mathrm{cov}_{x_\perp}\left(\sigma(\langle w_i, x\rangle + c_i), \sigma(\langle w_j, x\rangle + c_j)\right).\end{aligned}$$

Thus, returning to (1)–(2), we obtain

$$\mathcal{L}(a) = \frac{1}{2}\mathbb{E}_{x_\parallel}\left[\left(\varphi(U^\top x) - \sum_{i=1}^N a_i \overline{\sigma}_i(\langle w_{i\parallel}, x_\parallel\rangle + c_i)\right)^2\right] + \frac{1}{2}a^\top \Lambda a$$

$$= \frac{1}{2}\mathbb{E}_z\left[\left(\varphi(z) - \sum_{i=1}^N a_i \overline{\sigma}_i(\langle U^\top w_i, z\rangle + c_i)\right)^2\right] \tag{3}$$

$$+ \frac{1}{2}a^\top \Lambda a\,, \tag{4}$$

where $z = U^\top x$ and

$$\Lambda_{i,j} = \mathbb{E}_{x_\parallel}\left[\mathrm{cov}_{x_\perp}\left(\sigma(\langle w_i, x\rangle + c_i), \sigma(\langle w_j, x\rangle + c_j)\right)\right]\,.$$

The take-home message is that the high-dimensional regression problem in $x \in \mathbb{R}^d$ reduces to a lower dimensional regression problem in $z \in \mathbb{R}^k$ with an additional regularization term $a^\top \Lambda a$ and modified features. The non-linearities are smoothed and the feature vectors w_i are projected onto U. If $w_i \sim \mathcal{N}(0, \frac{1}{d}I_d)$, then $U^\top w_i \sim \mathcal{N}(0, \frac{1}{d}I_k)$. This gives small features: $\mathbb{E}\|U^\top w_i\|^2 = \frac{k}{d}$.

As a consequence, minimizing only the first term in (3)–(4), and taking the minimum norm solution, would lead to a minimum degree solution by Section 4. However, the second term, that controls the variance of the model in the orthogonal direction, actually has an important effect on the chosen minimizer. As we demonstrate in Example 5.3 below, in some cases it can break down the MDI bias.

Example 5.3. Consider the target function be $f(x) = 1$ with GOTU constraint $x_1 = 1$, and assume that the support of the training distribution contains a subset of the hyperplane $\{x \in \mathbb{R}^d \,|x_1 = 1\}$ of the form $\{1\} \times S_2 \times \cdots \times S_d$ with $|S_2|, \ldots, |S_d| \geqslant 3$. Then the MDI is given by the target function itself, but the random features model trained in sparse regime with $\sigma(x) = (1+x)^2$ converges to $f_{\mathrm{RF}}(x) = \frac{2}{5}x_1 + \frac{3}{5}$ (as $N \to \infty$ before $d \to \infty$). This shows that the random features model in general does not converge to the MDI in the sparse case, provided that the training distribution has strictly more than two inputs on each coordinate. Thus it is not possible to naively extend the results of (Abbe et al., 2023) beyond the hypercube $\{-1, 1\}^d$.

See the proof of Example 5.3 in Appendix C, and the simulation results in Figure 4. From Figure 4 we see that even for moderate values $d = 15$ and $N = 1024$, the model converges close to the asymptotic value.

Remark 5.4. Empirically we observed the lost of MDI property in this example for all polynomial activation functions that we checked, e.g. $(1+x)^2, (1+x)^2 - 1, x^2 + x, (1+x)^3, (1+x)^4$. Thus, we believe that it is a general property for polynomial activations rather than a degenerate case.

6. MDI for Data Embedded in Roots of Unity

We recall that $\mathbb{U}_n = \left\{\exp\left(i\frac{2\pi k}{n}\right) \mid k = 0, \ldots, n-1\right\} \subset \mathbb{C}$ denotes the n-roots of unity. Consider learning a target function $f : \mathbb{U}_n^d \to \mathbb{C}$ using a random feature model

$$f_{\mathrm{RF}}(a; x) = \frac{1}{\sqrt{N}}\sum_{i=1}^N a_i \phi_{w,b}(x)\,,$$

where the random features are defined as $\phi_{w,b}(x) = \sigma(\langle w, x\rangle + b)$. Compared to Section 3, this section takes the suitable generalization to the complex case: $a_i \in \mathbb{C}$, $b_i \in \mathbb{C}$ with distribution $(\Re b_i, \Im b_i) \sim \mathcal{N}\left(0, \frac{1}{d}I_2\right)$, $w_i \in \mathbb{C}^d$ with distribution $(\Re w_{i1}, \Im w_{i1}, \ldots, \Re w_{id}, \Im w_{id}) \sim \mathcal{N}\left(0, \frac{1}{d}I_{2d}\right)$, $\sigma : \mathbb{C} \to \mathbb{C}$ and $\langle w_i, x\rangle = \overline{w}_{i1}x_1 + \cdots + \overline{w}_{id}x_d$.

Let $\mathcal{U} \subset \mathbb{U}_n^d$ denote the subset of which f is unseen and denote

$$a^* = \operatorname*{arg\,min}_{a:\, f_{\mathrm{RF}}(a;x) = f(x),\, x \in \mathcal{U}^c} \|a\|$$

the minimum norm interpolant of f on the training domain. We recall that $\Pi_p(\mathbb{U}_n^d)$ denotes the set of complex polynomial functions of degree p on $\mathbb{U}_n^d$ (i.e. the set of functions on $\mathbb{U}_n^d$ that are the restriction to $\mathbb{U}_n^d$ of a multivariate complex polynomial on $\mathbb{C}^d$ of degree at most p).

Theorem 6.1. *Denote by p_* the minimum possible degree for a polynomial interpolator of f on the set $\mathcal{U}^c$. Then*

$$\lim_{d\to\infty}\lim_{N\to\infty} d(f_{\mathrm{RF}}(a^*, .), \Pi_p(\mathbb{U}_n^d)) = 0\,.$$

This result is proved in Appendix D.

7. Experiments

7.1. Experiments Setup

We run experiments[2] with the random features (RF) model and Transformer (Vaswani et al., 2017). For the RF model, we sample 65536 training points from the standard Gaussian distribution (except for the coordinates affected by the GOTU constraint, for which we simply hard code the required value) and train the model using Gradient Descent with line search (we refer to Appendix A.1 for the exact procedure). For convex functions with Lipschitz continuous gradient, this method provably converges to the global optimum and does not require the learning rate tuning. As for Transformer, we use AdamW optimizer (Loshchilov & Hutter, 2017) (without weight decay), and for each batch

[2] Code: https://github.com/DenisPushkin/GOTU-real-valued

we generate 256 random samples satisfying the GOTU constraint on the fly, imitating the access to the whole data on the seen domain. For the exact Transformer architecture we used, see Appendix A.2.

In both cases, we train the model on the data satisfying GOTU constraints and then evaluate on the full domain to capture its behavior on the unseen data. In case of the real-valued training domain, we evaluate the Hermite coefficients of the model. Note that the choice of Hermite polynomial basis is arbitrary, yet sufficient for our needs. Indeed, we are mainly interested in the polynomial degree of the function learnt by the model, which does not depend on the choice of polynomial basis.

When the training domain is a discrete grid, i.e. represented by $x \in \mathcal{X}^d$, where $\mathcal{X}$ is a finite set, we evaluate the model's coefficient considering it as a simple multivariate polynomial. It is justified by the fact that the set $\mathcal{B} = \{\prod_{i=1}^d x_i^{t_i} \mid 0 \leqslant t_i \leqslant |\mathcal{X}| - 1 \,\forall i\}$ of monomials with degree at most $|\mathcal{X}| - 1$ in each variable forms a basis of functions in $\mathcal{X}^d$. This result may be derived as a consequence of Combinatorial Nullstellensatz (Alon, 1999). Note that in a special case where $\mathcal{X} = \{\pm 1\}$, this basis produces the Fourier-Walsh basis of boolean functions, which was a central ingredient of MDI analysis in (Abbe et al., 2023).

7.2. Small Features Regime

First, we empirically confirm convergence to min-degree interpolator (MDI) for RF model in small features regime with polynomial activation (Theorem 4.2). We run two experiments: 1) $f(x) = 1$ with GOTU constraint $x_1 = 1$ (see Figure 1) and 2) $f(x) = x_2^2 + x_2 + 1$ with $x_1 = 1$ (Figure 2). In both cases, the target function f is itself an MDI, but in the second case the MDI is not unique: any function of the form $f(x) + (x_1 - 1)\Delta(x)$ with $deg(\Delta) \leqslant 1$ would be an MDI. As predicted by Theorem 4.2, the RF model converges to MDI in both cases. However, in the second experiment, the trained model depends on the variable x_1, while the target function does not. This shows that the random features model in small features regime does not always converge to "the simplest"[3] MDI.

7.3. Random Features Model with Standard Activations

Now, we examine the RF model with standard (non-polynomial) activations. First, we compare the sparse and the small features regimes on the target $f(x) = 1$ with GOTU constraint $x_1 = 1$. For both regimes, we use the same model architecture with $d = 15$ input dimension and $N = 1024$ random features and compare the same set of activation functions. You can see the result in Table 2 for small

[3]One possible formalization of "the simplest" MDI is a minimum degree-profile interpolator, defined in (Abbe et al., 2023).

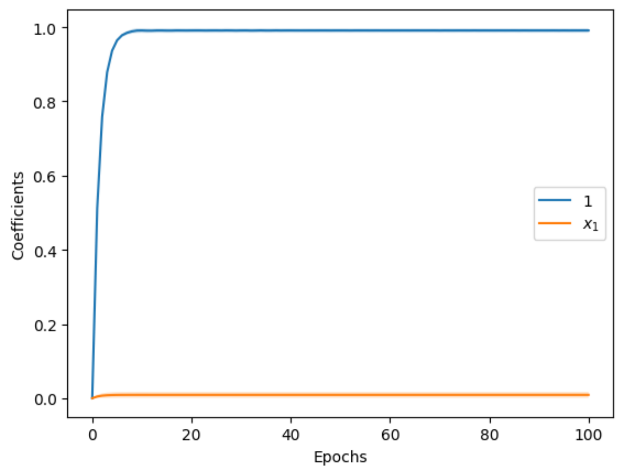

Figure 1. Training the random features model on $f(x) = 1$ with GOTU constraint $x_1 = 1$ in small features regime. Here, $d = 2$, $N = 256$, $\varepsilon = (0.05)^2$, and $\sigma(x) = (1 + x)^2$.

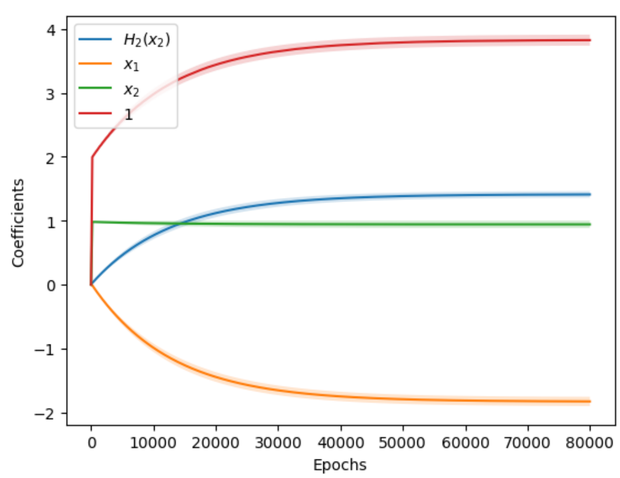

Figure 2. Training the random features model on $f(x) = x_2^2 + x_2 + 1$ with GOTU constraint $x_1 = 1$ in small features regime. Here, $d = 2$, $N = 16384$, $\varepsilon = (0.05)^2$ and $\sigma(x) = (1 + x)^2$. The model converged to MDI, but not "the simplest one", since it depends on x_1.

Table 1. Training the random features model on $f(x) = x_1x_2$, $x \in \mathbb{R}^d$ with GOTU constraint $(x_1 - 1)(x_2 - 1) = 0$ in sparse regime. Here, $d = 15$ and $N = 40000$. The MDI is given by $x_1 + x_2 - 1$.

ACTIVATION	1	x_1	x_2	x_1x_2
RELU	-0.952 ± 0.002	0.955 ± 0.005	0.955 ± 0.004	0.042 ± 0.009
SHIFTED RELU	-0.957 ± 0.002	0.958 ± 0.004	0.958 ± 0.005	0.039 ± 0.009
SIGMOID	-1.013 ± 0.003	0.996 ± 0.003	0.996 ± 0.006	-0.001 ± 0.009
SOFTPLUS	-0.975 ± 0.003	0.978 ± 0.004	0.977 ± 0.006	0.022 ± 0.010

Table 2. Training the random features model on $f(x) = 1$, $x \in \mathbb{R}^d$ with GOTU constraint $x_1 = 1$ in small features regime with $\varepsilon = (0.03)^2$. Here, $d = 15$ and $N = 1024$.

ACTIVATION	1	x_1
$(1+x)^2$	0.997 ± 0.002	0.001 ± 0.003
RELU	0.564 ± 0.009	0.430 ± 0.010
SHIFTED RELU	1.000 ± 0.000	-0.001 ± 0.003
SIGMOID	1.000 ± 0.000	-0.001 ± 0.003
SOFTPLUS	1.000 ± 0.001	-0.001 ± 0.003

Table 3. Training the random features model on $f(x) = 1$, $x \in \mathbb{R}^d$ with GOTU constraint $x_1 = 1$ in sparse regime. Here, $d = 15$ and $N = 1024$.

ACTIVATION	1	x_1
$(1+x)^2$	0.624 ± 0.017	0.374 ± 0.017
RELU	0.564 ± 0.009	0.431 ± 0.011
SHIFTED RELU	0.782 ± 0.009	0.217 ± 0.011
SIGMOID	0.992 ± 0.003	0.007 ± 0.002
SOFTPLUS	0.789 ± 0.010	0.208 ± 0.012

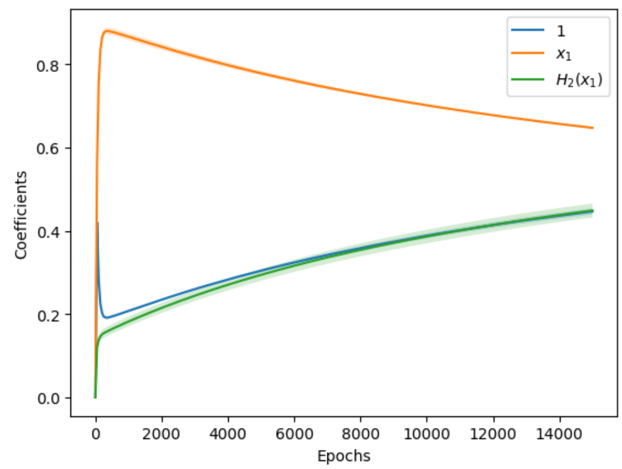

Figure 3. Training the random features model on $f(x) = 1$ with GOTU constraint $x_1 = 1$ in sparse regime with $\sigma(x) = (1 + x)^4$ activation. Here, $d = 15$, $N = 3 \cdot 10^5$, and $H_2(x_1)$ denotes the normalized second degree Hermite polynomial. The MDI is a constant function 1, but the model learns the quadratic function.

features regime and Table 3 for sparse regime. We see that in sparse regime, the RF model in general learns a linear interpolator instead of the constant one (the only exception is Sigmoid activation). For small features regime, the RF model converges to the constant interpolator for all activations, except for ReLU. We conjecture that this happens because $\text{ReLU}(0) = 0$, which breaks the Assumption 4.1, used in the Theorem 4.2 (note that this assumption was stated only for polynomial activations). In contrast, with Shifted ReLU activation, given by Shifted $\text{ReLU}(x) = \text{ReLU}(x) - 1$, the convergence to MDI is restored. Hence, we conjecture that Assumption 4.1 is also a prerequisite for non-polynomial activations to guarantee the convergence to MDI in small features regime.

Next, we train RF model in sparse regime on $f(x) = x_1x_2$ with GOTU constraint $(x_1 - 1)(x_2 - 1) = 0$ (see Table 1), where the MDI is given by $x_1 + x_2 - 1$. In this example, the RF model converges close to MDI for all activations we tried, which shows that, for some target functions, the MDI bias may also holds for the RF model in sparse regime.

Finally, Example 5.3 illustrates that the RF model in sparse regime with polynomial activation generally does not learn the MDI. But how far can it depart from the MDI, e.g. can the degree of the trained model exceed the minimum degree of interpolator by more than one? In Figure 3 we demonstrate that the RF model with $\sigma(x) = (1 + x)^4$ activation trained on $f(x) = 1$ with GOTU constraint $x_1 = 1$ learns the quadratic function. It shows that the RF model in sparse regime can exceed the MDI by more than one degree. We conjecture that this gap can be arbitrarily large as we increase the degree of the polynomial activation function.

7.4. Transformer and Random Features with Discrete Input

Finally, we consider the input variable x from the discrete space. It allows us to apply Transformer model in our experiments.

For Transformer, we run the following experiments: 1) $f(x) = 1$, $x \in \{-2, -1, 0, 1, 2\}^d$ with GOTU constraint

$x_1 = 1$ in dimension $d = 15$ (see Figure 5) and 2) $f(x) = x_1x_2$, $x \in \{-1, 0, 1\}^d$ with $(x_1 - 1)(x_2 - 1) = 0$ and $d = 15$ (Figure 6). In the first experiment, the MDI is given by a target function $f(x) = 1$ itself, and Transformer indeed learns a constant function and neglects the constrained variable x_1. In the second experiment, the MDI is given by a linear function $f(x) = x_1 + x_2 - 1$. In this case, the Transformer's behavior depends on the learning rate. With a moderate learning rate of 10^{-4} (Figure 6, left), Transformer's coefficients are noisy at the first half of the training, but then sharply stabilize and converge to the interpolator[4] $f_{int}(x) = \frac{1}{2}(x_1 + x_2 - x_1^2 + x_1x_2 - x_2^2 + x_1^2x_2^2)$ (see the exact coefficients in Table 4). This shows that Transformer consistently learns the interpolator of degree 4 instead of the linear MDI. We also repeat this experiment with 10^{-5} learning rate - the same one which leads to the min-degree interpolator for boolean functions in (Abbe et al., 2023) (Figure 6, right). In this case, Transformer's coefficients did not converge even after $6 \cdot 10^6$ iterations. However, the trajectory suggests that the coefficient of x_1x_2 is non-negligible, which means that Transformer learns at least a quadratic function. Moreover, the coefficient x_1x_2 is likely to dominate all other coefficients, implying that the learnt function is not an MDI even in a leaky sense (i.e. the high-degree monomials are not dominated by the low-degree alternatives).

Note the crucial difference with the boolean case, where Transformer converges to MDI when trained on the same target function with 10^{-5} learning rate (Abbe et al., 2023).[5] This shows that min-degree bias for Transformer does not generalize beyond the boolean domain.

We also train RF model on $f(x) = 1$, $x \in \{-2, -1, 0, 1, 2\}^d$ with GOTU constraint $(x_1 - 1)(x_2 - 1) = 0$ in dimension $d = 15$, using $\sigma(x) = (1 + x)^2$ activation (see Figure 4). We observe that the RF model learns a linear function, while the MDI is given by a constant. It confirms the statement of Example 5.3 that even for discrete domains, the RF model in sparse regime with polynomial activation does not converge to MDI.

[4]Of course, it's just a hypothesis that Transformer converges to this exact function. In the experiments, the final coefficients of the Transformer are very close to $\pm 1/2$, but never equal to it.

[5]The other distinction between our experiments and the ones made by (Abbe et al., 2023) is that the latter stops the training when the loss reaches a low enough threshold, while we train the model longer until its coefficients are well stabilized. It may happen that (leaky) MDI bias is stronger when lower learning rates or early stopping is used; we leave this hypothesis, as well as the evolution of the MDI on long training past a 'low' threshold for future research.

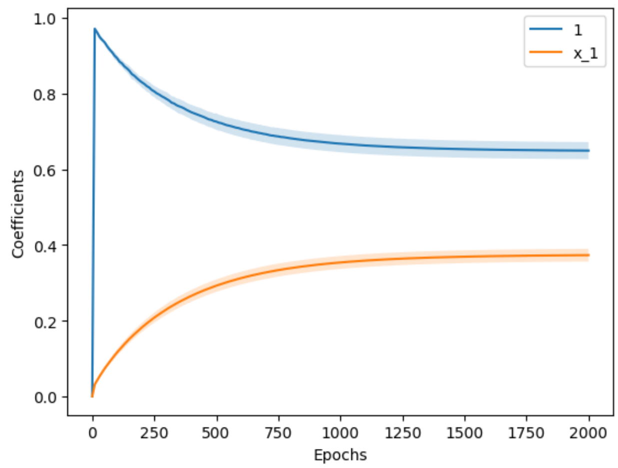

Figure 4. Training the random features model on $f(x) = 1$, $x \in \{-2, -1, 0, 1, 2\}^d$ with GOTU constraint $x_1 = 1$ and $\sigma(x) = (1 + x)^2$ activation. Here, $d = 15$, $N = 1024$. While MDI is given by a constant function, the model learns a linear interpolator.

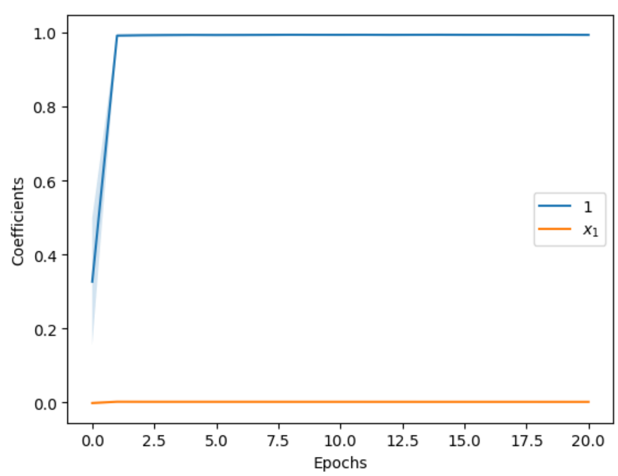

Figure 5. Training Transformer on $f(x) = 1$, $x \in \{-2, -1, 0, 1, 2\}^d$ with GOTU constraint $x_1 = 1$ using AdamW optimizer with 10^{-4} learning rate. Here, $d = 15$.

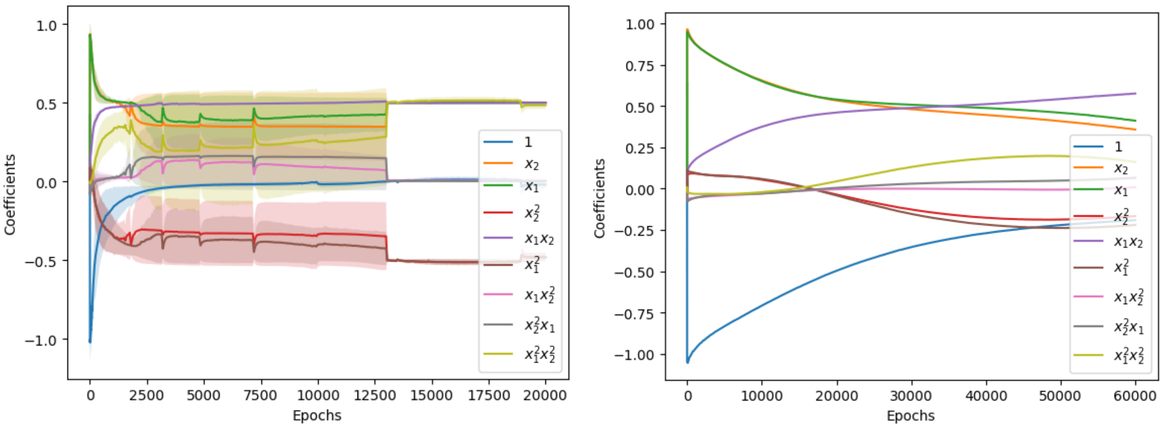

Figure 6. Training Transformer on $f(x) = x_1x_2, x \in \{-1,0,1\}^d$ with GOTU constraint $(x_1-1)(x_2-1) = 0$ in dimension $d = 15$ using AdamW optimizer. The MDI is given by $x_1 + x_2 - 1$. We used the learning rate 10^{-4} for the left plot, and 10^{-5} for the right one.

Table 4. Final coefficients of the trained Transformer on $f(x) = x_1x_2, x \in \{-1,0,1\}^d$ with GOTU constraint $(x_1-1)(x_2-1) = 0$ in dimension $d = 15$. Here, we used AdamW optimizer with 10^{-4} learning rate.

Monomial	Coefficient
1	-0.020 ± 0.028
x_1	0.499 ± 0.000
x_2	0.501 ± 0.001
x_1^2	-0.480 ± 0.027
x_1x_2	0.500 ± 0.004
x_2^2	-0.481 ± 0.027
$x_1^2x_2$	0.002 ± 0.005
$x_2^2x_1$	0.005 ± 0.003
$x_1^2x_2^2$	0.484 ± 0.024

8. Additional Related Literature

This paper is a generalization and extension of (Abbe et al., 2023).

Out-of-distribution generalization is a critical aspect of machine learning that has been studied both in theory (Ben-David et al., 2006; Mansour et al., 2009; Redko et al., 2020) and in practice (Gulrajani & Lopez-Paz, 2020; Miller et al., 2021; Wiles et al., 2022). Our work considers an extreme case of distribution shift in which part of the domain is entirely unseen during the training. OOD generalization and the ability to extrapolate have also been used as proxies for measuring the reasoning capabilities of neural networks (Saxton et al., 2019; Zhang et al., 2021; Csordás et al., 2021; Zhang et al., 2022) as these models are prone to memorization of training samples (Carlini et al., 2019; Feldman & Zhang, 2020; Kandpal et al., 2022; Carlini et al., 2022; Zhang et al., 2021) or learning undesirable shortcuts (Zhang et al., 2022). A special case is length generalization (Zaremba & Sutskever, 2014; Lake & Baroni, 2018; Hupkes et al., 2020; Zhang et al., 2022; Anil et al., 2022), i.e., generalization to the input lengths beyond what is seen during the training.

It has been shown that training with gradient descent imposes particular implicit regularization on the solutions found by the models such as sparsity (Moroshko et al., 2020), norm minimization (Bartlett et al., 2021), and margin maximization (in linear classification setting) (Soudry et al., 2017). This implicit regularization (or implicit bias) of neural networks trained with gradient-based algorithms has been used to explain the generalization of (often over-parametrized) models (Bartlett et al., 2021). These results depend on the optimizer (Gunasekar et al., 2018a) and model (Gunasekar et al., 2018b) and are usually proven for simple models such as linear models (Soudry et al., 2017; Yun et al., 2020; Jacot et al., 2021) including diagonal linear neural networks (Gunasekar et al., 2018b; Moroshko et al., 2020). Our result for the random feature model builds upon the implicit bias toward solutions with minimum norm (Bartlett et al., 2021). Related to us is also the spectral bias (Xu et al., 2019; Rahaman et al., 2019) stating that neural networks, when learning a function in continuous settings, capture the lower frequency components faster (note that degree in Boolean functions plays a similar role to the frequency).

9. Conclusion

This paper shows that the min-degree bias in the non-Boolean case is mitigated by various phenomena. One setting admits a clear min-degree bias, i.e., the roots of unities. Moreover, Transformer may still admit some leaky min-degree bias, and it remains open to understand what else drives the bias of Transformers; e.g., does the produced output alphabet or influence of variables play a role?

Impact Statement

This paper presents work whose goal is to advance the field of Machine Learning. There are many potential societal consequences of our work, none which we feel must be specifically highlighted here.

References

Abbe, E., Bengio, S., Lotfi, A., and Rizk, K. Generalization on the unseen, logic reasoning and degree curriculum. In *ICML*, 2023. URL https://arxiv.org/abs/2301.13105.

Alon, N. Combinatorial nullstellensatz. *Combinatorics, Probability and Computing*, 8(1-2):7–29, 1999.

Anil, C., Wu, Y., Andreassen, A., Lewkowycz, A., Misra, V., Ramasesh, V., Slone, A., Gur-Ari, G., Dyer, E., and Neyshabur, B. Exploring length generalization in large language models. *arXiv preprint arXiv:2207.04901*, 2022.

Bakhtin, A., van der Maaten, L., Johnson, J., Gustafson, L., and Girshick, R. Phyre: A new benchmark for physical reasoning. *Advances in Neural Information Processing Systems*, 32, 2019.

Bartlett, P. L., Montanari, A., and Rakhlin, A. Deep learning: a statistical viewpoint. *Acta numerica*, 30:87–201, 2021.

Ben-David, S., Blitzer, J., Crammer, K., and Pereira, F. Analysis of representations for domain adaptation. *Advances in neural information processing systems*, 19, 2006.

Carlini, N., Liu, C., Erlingsson, Ú., Kos, J., and Song, D. The secret sharer: Evaluating and testing unintended memorization in neural networks. In *USENIX Security Symposium*, volume 267, 2019.

Carlini, N., Ippolito, D., Jagielski, M., Lee, K., Tramer, F., and Zhang, C. Quantifying memorization across neural language models. *arXiv preprint arXiv:2202.07646*, 2022.

Csordás, R., Irie, K., and Schmidhuber, J. The devil is in the detail: Simple tricks improve systematic generalization of transformers. *arXiv preprint arXiv:2108.12284*, 2021.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

Feldman, V. and Zhang, C. What neural networks memorize and why: Discovering the long tail via influence estimation. *Advances in Neural Information Processing Systems*, 33:2881–2891, 2020.

Gulrajani, I. and Lopez-Paz, D. In search of lost domain generalization. *arXiv preprint arXiv:2007.01434*, 2020.

Gunasekar, S., Lee, J., Soudry, D., and Srebro, N. Characterizing implicit bias in terms of optimization geometry, 2018a. URL https://arxiv.org/abs/1802.08246.

Gunasekar, S., Lee, J., Soudry, D., and Srebro, N. Implicit bias of gradient descent on linear convolutional networks, 2018b. URL https://arxiv.org/abs/1806.00468.

Hendrycks, D. and Gimpel, K. Gaussian error linear units (gelus). *arXiv preprint arXiv:1606.08415*, 2016.

Hupkes, D., Dankers, V., Mul, M., and Bruni, E. Compositionality decomposed: How do neural networks generalise? *Journal of Artificial Intelligence Research*, 67: 757–795, 2020.

Jacot, A., Ged, F., Şimşek, B., Hongler, C., and Gabriel, F. Saddle-to-saddle dynamics in deep linear networks: Small initialization training, symmetry, and sparsity. *arXiv preprint arXiv:2106.15933*, 2021.

Johnson, J., Hariharan, B., Van Der Maaten, L., Fei-Fei, L., Lawrence Zitnick, C., and Girshick, R. Clevr: A diagnostic dataset for compositional language and elementary visual reasoning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2901–2910, 2017.

Kandpal, N., Wallace, E., and Raffel, C. Deduplicating training data mitigates privacy risks in language models. In *International Conference on Machine Learning*, pp. 10697–10707. PMLR, 2022.

Lake, B. and Baroni, M. Generalization without systematicity: On the compositional skills of sequence-to-sequence recurrent networks. In *International conference on machine learning*, pp. 2873–2882. PMLR, 2018.

Lewkowycz, A., Andreassen, A., Dohan, D., Dyer, E., Michalewski, H., Ramasesh, V., Slone, A., Anil, C., Schlag, I., Gutman-Solo, T., et al. Solving quantitative reasoning problems with language models. *arXiv preprint arXiv:2206.14858*, 2022.

Loshchilov, I. and Hutter, F. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.

Mahdavi, S., Swersky, K., Kipf, T., Hashemi, M., Thrampoulidis, C., and Liao, R. Towards better out-of-distribution generalization of neural algorithmic reasoning tasks. *ArXiv*, 2211.00692, 2022. URL https://arxiv.org/abs/2211.00692.

Mansour, Y., Mohri, M., and Rostamizadeh, A. Domain adaptation: Learning bounds and algorithms. *arXiv preprint arXiv:0902.3430*, 2009.

Miller, J. P., Taori, R., Raghunathan, A., Sagawa, S., Koh, P. W., Shankar, V., Liang, P., Carmon, Y., and Schmidt, L. Accuracy on the line: on the strong correlation between out-of-distribution and in-distribution generalization. In *International Conference on Machine Learning*, pp. 7721–7735. PMLR, 2021.

Moroshko, E., Gunasekar, S., Woodworth, B., Lee, J. D., Srebro, N., and Soudry, D. Implicit bias in deep linear classification: Initialization scale vs training accuracy, 2020. URL https://arxiv.org/abs/2007.06738.

Rahaman, N., Baratin, A., Arpit, D., Draxler, F., Lin, M., Hamprecht, F., Bengio, Y., and Courville, A. On the spectral bias of neural networks. In *International Conference on Machine Learning*, pp. 5301–5310. PMLR, 2019.

Redko, I., Morvant, E., Habrard, A., Sebban, M., and Bennani, Y. A survey on domain adaptation theory: learning bounds and theoretical guarantees. *arXiv preprint arXiv:2004.11829*, 2020.

Saxton, D., Grefenstette, E., Hill, F., and Kohli, P. Analysing mathematical reasoning abilities of neural models. *arXiv preprint arXiv:1904.01557*, 2019.

Soudry, D., Hoffer, E., Nacson, M. S., Gunasekar, S., and Srebro, N. The implicit bias of gradient descent on separable data, 2017. URL https://arxiv.org/abs/1710.10345.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

Veličković, P., Badia, A. P., Budden, D., Pascanu, R., Banino, A., Dashevskiy, M., Hadsell, R., and Blundell, C. The clrs algorithmic reasoning benchmark. *arXiv preprint arXiv:2205.15659*, 2022.

Wiles, O., Gowal, S., Stimberg, F., Rebuffi, S.-A., Ktena, I., Dvijotham, K. D., and Cemgil, A. T. A fine-grained analysis on distribution shift. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=Dl4LetuLdyK.

Xu, Z.-Q. J., Zhang, Y., Luo, T., Xiao, Y., and Ma, Z. Frequency principle: Fourier analysis sheds light on deep neural networks, 2019.

Yun, C., Krishnan, S., and Mobahi, H. A unifying view on implicit bias in training linear neural networks. *arXiv preprint arXiv:2010.02501*, 2020.

Zaremba, W. and Sutskever, I. Learning to execute. *arXiv preprint arXiv:1410.4615*, 2014.

Zhang, C., Raghu, M., Kleinberg, J. M., and Bengio, S. Pointer value retrieval: A new benchmark for understanding the limits of neural network generalization. *ArXiv*, abs/2107.12580, 2021.

Zhang, Y., Backurs, A., Bubeck, S., Eldan, R., Gunasekar, S., and Wagner, T. Unveiling transformers with lego: a synthetic reasoning task. *arXiv preprint arXiv:2206.04301*, 2022.

A. Experiments Details

A.1. Gradient Descent with Line Search

Algorithm 1 Gradient Descent with Line Search

Input: data point x_0, Liphitz constant estimator $L_0 = 1$
for $n = 0, 1, \dots$ **do**
  $x_{n+1} = x_n - \frac{1}{L_n} \nabla f(x_n)$
  while not $f(x_{n+1}) \leqslant f(x_n) - \frac{1}{2L_n} \|\nabla f(x_n)\|^2$ **do**
    $L_n := 2L_n$
    $x_{n+1} = x_n - \frac{1}{L_n} \nabla f(x_n)$
  end while
  $L_{n+1} = L_n / 2$
end for

A.2. Transformer Architecture

For Transformer, we use the encoder-only architecture from the Vision Transformer model (Dosovitskiy et al., 2020). This model consists of 12 layers, each of them formed by multi-head self-attention block with 6 heads followed by Feed-Forward block. Following standard practices, there is a layer normalization before each self-attention and Feed-Forward blocks. The model uses decoupled weights, i.e. there is no parameters sharing between the layers or the attention heads. For each input sequence, the model prepends a special classification token at the beginning of the sequence. Then it encodes each token (which comes from the discrete alphabet) using the input embedding layer and adds it to the trainable positional embedding. We keep the embedding dimension equal to 64 both at the beginning of the model and between the model blocks. The Feed-Forward module is represented by a 2-layers MLP with hidden dimension 64 and GELU activation (Hendrycks & Gimpel, 2016). To get the final prediction, the model extracts the final classification token embedding, and feeds it through the layer normalization followed by a linear layer with a single output.

B. Proof of Theorem 4.2

Reminder on the Hermite decomposition. Let H_t denote the Hermite polynomial of degree t, using the probabilist convention, and normalized such that $\{H_t,\, t \geqslant 0\}$ is an orthonormal basis of $L^2(\mathbb{R}, \gamma_1)$. (We recall that γ_d denotes the standard Gaussian measure in dimension d.) Said differently, if Z is a univariate standard normal random variable, we assume that $\mathbb{E}\left[H_s(Z)H_t(Z)\right] = \mathbb{1}_{s=t}$. Further, we define the multivariate Hermite polynomials as $\chi_T(x) = \prod_{i=1}^d H_{t_i}(x_i)$, where $T = (t_1, \dots, t_d) \in \mathbb{N}^d$. Note that $\deg(\chi_T(x)) = |T| = T_1 + \dots + T_d$. The set of functions $\{\chi_T(x), T \in \mathbb{N}^d\}$ forms an orthonormal basis of $L^2(\mathbb{R}^d, \gamma_d)$.

Recall that $\Pi_p(\mathbb{R}^d)$ denotes the set of polynomials of degree at most p on $\mathbb{R}^d$. Any $h \in \Pi_p(\mathbb{R}^d)$ admits a Hermite decomposition of the form $h(x) = \sum_{T \in \mathbb{N}^d_{\leqslant p}} \widehat{h}(T)\chi_T(x)$, where $\mathbb{N}^d_{\leqslant p} = \{T \in \mathbb{N}^d \mid |T| \leqslant p\}$ and $\widehat{h}(T) = \mathbb{E}_Z[h(Z)\chi_T(Z)]$, $Z \sim \mathcal{N}(0, I_d)$.

We now turn to the proof of the theorem.

Lemma B.1. *If $\sigma \in \Pi_p(\mathbb{R})$, then w.h.p. we have that for any large enough N,*

$$\mathrm{im}(f_{\mathrm{RF}}) = \Pi_p(\mathbb{R}^d)\,.$$

Proof. Since $\sigma \in \Pi_p(\mathbb{R})$, we have that $\forall i \in [N]$: $\phi_{w_i,b_i}(x) = \sigma(\langle w_i, x\rangle + b_i) \in \Pi_p(\mathbb{R}^d)$. Thus, $f_{\mathrm{RF}}(a; x) = \frac{1}{\sqrt{N}} \sum_{i=1}^N a_i \phi_{w_i,b_i}(x) \in \Pi_p(\mathbb{R}^d)\ \forall a \in \mathbb{R}^N$, which shows that $\mathrm{im}(f_{\mathrm{RF}}) \subseteq \Pi_p(\mathbb{R}^d)$.

It remains to show that $\mathrm{im}(f_{\mathrm{RF}}) \supseteq \Pi_p(\mathbb{R}^d)$. Let γ_d denote the standard Gaussian measure in dimension d. We define the operator $M : \Pi_p(\mathbb{R}^d) \to \Pi_p(\mathbb{R}^d)$ by the formula

$$M(f) = \mathbb{E}_{w,b}\left[\langle f, \sigma(\langle w, x\rangle + b)\rangle_{L^2(\gamma_d)}\, \sigma(\langle w, x\rangle + b)\right]\,, \qquad f \in \Pi_d(\mathbb{R}^d)\,.$$

Similarly, we define the empirical version $M_N : \Pi_p(\mathbb{R}^d) \to \Pi_p(\mathbb{R}^d)$ by the formula

$$M_N(f) = \frac{1}{N}\sum_{i=1}^{N} \langle f, \sigma(\langle w_i, x\rangle + b_i)\rangle_{L^2(\gamma_d)} \sigma(\langle w_i, x\rangle + b_i)\,, \qquad f \in \Pi_d(\mathbb{R}^d)\,.$$

The operators M and M_N are positive definite over $(\Pi_p(\mathbb{R}^d), \langle ., .\rangle_{L^2(\gamma_d)})$.

Assume that M is positive definite. By the law of large numbers, $M_N \xrightarrow[N\to\infty]{} M$ almost surely. As the set of positive definite matrices is an open set, this implies that for all $\eta > 0$, there exists $N_0 \in \mathbb{N}$ such that for all $N \geqslant N_0$, $\Pr(M_N \text{ is positive definite}) \geqslant 1 - \eta$.

When M_N is positive definite, then $\mathrm{im}(M_N) = \Pi_p(\mathbb{R}^d)$. As $\mathrm{im}(M_N) \subset \mathrm{im}(f_{\mathrm{RF}}) \subset \Pi_p(\mathbb{R}^d)$, this enables to conclude that $\mathrm{im}(f_{\mathrm{RF}}) = \Pi_p(\mathbb{R}^d)$.

We thus now need to prove that M is positive definite. Consider f such that $\langle f, M(f)\rangle_{L^2(\gamma_d)} = 0$. We prove that $f = 0$.

We have that $\mathbb{E}_{w,b}\left[\langle f, \sigma(\langle w, x\rangle + b)\rangle^2_{L^2(\gamma_d)}\right] = 0$. Thus $\langle f, \sigma(\langle w, x\rangle + b)\rangle_{L^2(\gamma_d)} = 0$ almost surely. As this expression is a multivariate polynomial in w and b, this implies that actually $\langle f, \sigma(\langle w, x\rangle + b)\rangle_{L^2(\gamma_d)} = 0$ for all $w \in \mathbb{R}^d$ and $b \in \mathbb{R}$. In particular, if $\lambda \geqslant 0$, we have $\langle f, \sigma(\lambda\langle w, x\rangle)\rangle_{L^2(\gamma_d)} = 0$.

We use a Taylor expansion for σ:

$$\sigma(y) = \sum_{k=0}^{p} \frac{\sigma^{(k)}(0)}{k!} y^k\,.$$

As a consequence,

$$0 = \langle f, \sigma(\lambda\langle w, x\rangle)\rangle_{L^2(\gamma_d)} = \sum_{k=0}^{p}\left\langle f, \frac{\sigma^{(k)}(0)}{k!}\left(\lambda\langle w, x\rangle\right)^k\right\rangle_{L^2(\gamma_d)} = \sum_{k=0}^{p}\frac{\sigma^{(k)}(0)}{k!}\lambda^k \left\langle f, \langle w, x\rangle^k\right\rangle_{L^2(\gamma_d)}\,.$$

Identifying powers of λ in this expression, and using that $\sigma^{(k)}(0) \neq 0$ for all k, we have that

$$\left\langle f, \langle w, x\rangle^k\right\rangle_{L^2(\gamma_d)}\,, \qquad k = 0, \dots, p\,.$$

To conclude, we only need to prove that the set of functions $\langle w, x\rangle^k$ for $w \in \mathbb{R}^d$ and $k = 0, \dots, p$ spans $\Pi_p(\mathbb{R}^d)$.

Consider w such that $\|w\| = 1$. We decompose f into multivariate Hermite polynomials: $f(x) = \sum_{|\beta|\leqslant p}\widehat{f}(\beta)h_\beta(x)$. Then

$$0 = \langle f, h_k(\langle w, x\rangle)\rangle_{L^2(\gamma_d)} = \sum_{|\beta|\leqslant p}\widehat{f}(\beta)\,\langle h_\beta(x), h_k(\langle w, x\rangle)\rangle_{L^2(\gamma_d)} = \sum_{|\beta|\leqslant p}\widehat{f}(\beta)\binom{k}{\beta}^{1/2} w_1^{\beta_1}\cdots w_d^{\beta_d}\,.$$

The last quantity is a multivariate polynomial in w, which is zero on the unit sphere. It thus need to be identically zero. Thus for all β, $\widehat{f}(\beta) = 0$. Thus $f = 0$. This concludes the proof. $\square$

Lemma B.2. *Assume that $A_n \to A$ as $n \to \infty$, where $(A_n)_{n=1}^{\infty}$, A are positive-definite matrices in $\mathbb{R}^{d\times d}$ and let $\mathcal{S}$ be any affine subspace of $\mathbb{R}^d$. Then*

$$\mathit{argmin}_{x\in\mathcal{S}} x^\top A_n x \to \mathit{argmin}_{x\in\mathcal{S}} x^\top A x$$

Proof. Let us introduce the following notations:

$$y_n = \mathrm{argmin}_{x\in\mathcal{S}} x^\top A_n x$$
$$y = \mathrm{argmin}_{x\in\mathcal{S}} x^\top A x$$
$$\rho = \min_{x\in\mathcal{S}} x^\top A x = y^\top A y$$

First, let us show that $\|y_n\|_2$ is uniformly bounded over n. From $A_n \to A$ we get that $\lambda_{min}(A_n) \to \lambda_{min}(A)$ and $y^\top A_n y \to y^\top Ay$. Thus, $\exists n_0 \in \mathbb{N}$ s.t. $\forall n \geqslant n_0$ both of the following holds: $\lambda_{min}(A_n) \geqslant \lambda_{min}(A)/2 > 0$ and $|y^\top A_n y - y^\top Ay| < 1$. We claim that for any such n it holds that $\|y_n\|_2 \leqslant C \overset{def}{=} \sqrt{\frac{2(\rho+1)}{\lambda_{min}(A)}}$. Indeed, consider any x such that $\|x\|_2 > C$. Then

$$x^\top A_n x \geqslant \lambda_{min}(A_n)\|x\|_2^2 \geqslant \frac{\lambda_{min}(A)}{2}\|x\|_2^2 > \frac{\lambda_{min}(A)}{2}\frac{2(\rho+1)}{\lambda_{min}(A)} = \rho + 1$$

On the other hand, $y^\top A_n y < y^\top Ay + 1 = \rho + 1$. Hence, $y^\top A_n y < x^\top A_n x$, which shows that none of such x can be the minimizer. This concludes $\|y\|_2 \leqslant C$. As a side effect, from $y^\top A_n y < x^\top A_n x \; \forall x : \|x\|_2 > C$ we get $\|y\|_2 \leqslant C$.

To complete the proof, it is enough to show that y is a unique partial limit of the sequence (y_n). Define $\mathcal{S}' = \{x \in \mathcal{S} : \|x\|_2 \leqslant C\}$. As we have proved above, $y, y_n \in \mathcal{S}'$ for large enough n. Let us show that the functions $x \to x^\top A_n x$ converge to $x \to x^\top Ax$ uniformly over $x \in \mathcal{S}'$. Indeed,

$$|x^\top A_n x - x^\top Ax| \leqslant \|x\|_2 \cdot \|(A_n - A)x\|_2 \leqslant \|A_n - A\|_2 \cdot \|x\|_2^2$$

and the last term converges to zero uniformly whenever $\|x\|_2$ is uniformly bounded.

Now let l be any partial limit of y_n, that is $\exists(y_{n_k})$: $y_{n_k} \to l$. We want to show that $l = y$. Let us fix any $\varepsilon > 0$. From the uniform convergence we know that $\exists n_0 \in \mathbb{N}$ s.t. $|x^\top A_n x - x^\top Ax| < \varepsilon \; \forall x \in \mathcal{S}', \forall n \geqslant n_0$. Recall that $y, y_n \in \mathcal{S}'$ for large enough n, which means all the elements of (y_{n_k}) belong to $\mathcal{S}'$ starting from some k_0. For these elements, we can estimate

$$\rho = y^\top Ay \leqslant y_{n_k}^\top A y_{n_k} \leqslant y_{n_k}^\top A_{n_k} y_{n_k} + \varepsilon \leqslant y^\top A_{n_k} y + \varepsilon \leqslant y^\top Ay + 2\varepsilon = \rho + 2\varepsilon$$

where the first inequality comes from $y = \text{argmin}_{x\in\mathcal{S}} x^\top Ax$, and the third - from $y_{n_k} = \text{argmin}_{x\in\mathcal{S}} x^\top A_{n_k} x$. Thus, $\forall\varepsilon > 0$ we get

$$\rho \leqslant y_{n_k}^\top A y_{n_k} \leqslant \rho + 2\varepsilon \quad \forall k \geqslant k_0(\varepsilon)$$

which shows that $y_{n_k}^\top A y_{n_k} \to \rho = \min_{x\in\mathcal{S}} x^\top Ax$. Taking into account that the function $x \to x^\top Ax$ is strongly convex (since A is positive-definite), we conclude that $y_{n_k} \to y$. Hence, the only partial limit of (y_n) is y, which proves that $y_n \to y$. $\square$

Lemma B.3. *For any basis monomial $\chi_T(x)$ and any non-negative integer $k < |T|$ we have:*

$$\mathbb{E}_{x\sim\mathcal{N}(0,I_d)}[(\langle w, x\rangle + b)^k \chi_T(x)] = 0$$

Proof. Since the term $(\langle w, x\rangle + b)^k$ is a polynomial of degree k in x, in Hermite basis it only contains Hermite polynomials of degree at most k. The rest comes from the orthogonality of Hermite polynomials. $\square$

Lemma B.4. *For any non-negative integer k we have:*

$$\mathbb{E}_{x\sim\mathcal{N}(0,I_d)}[|(\langle w, x\rangle + b)^k \chi_T(x)|] \leqslant \varepsilon^{k/2} poly_T(\bar{w}, \bar{b})$$

where $poly_T(\bar{w}, \bar{b})$ is some polynomial in $\bar{w}_1, \ldots, \bar{w}_d, \bar{b}$ which depends on T. Here, $\bar{w} = \varepsilon^{-1/2} w$, $\bar{b} = \varepsilon^{-1/2} b$.

Proof.

$$\begin{aligned}
\mathbb{E}_x[|(\langle w, x\rangle + b)^k \chi_T(x)|] &= \varepsilon^{k/2}\mathbb{E}_x[|(\langle \bar{w}, x\rangle + \bar{b})^k \chi_T(x)|] \\
&= \varepsilon^{k/2}\mathbb{E}_x\left[\left|\sum_{\alpha_1+\ldots+\alpha_{d+1}=k}\binom{k}{\alpha_1 \ldots \alpha_{d+1}}(\bar{w}_1 x_1)^{\alpha_1}\ldots(\bar{w}_d x_d)^{\alpha_d}\bar{b}^{\alpha_{d+1}}\chi_T(x)\right|\right] \\
&\leqslant \varepsilon^{k/2}\sum_{\alpha_1+\ldots+\alpha_{d+1}=k}\binom{k}{\alpha_1 \ldots \alpha_{d+1}}|\bar{w}_1^{\alpha_1}\ldots\bar{w}_d^{\alpha_d}\bar{b}^{\alpha_{d+1}}| \cdot \mathbb{E}_x[|x_1^{\alpha_1}\ldots x_d^{\alpha_d}\chi_T(x)|] \\
&\leqslant \varepsilon^{k/2}\sum_{\alpha_1+\ldots+\alpha_{d+1}=k}\binom{k}{\alpha_1 \ldots \alpha_{d+1}}\left(\frac{\bar{w}_1^2+1}{2}\right)^{\alpha_1}\ldots\left(\frac{\bar{w}_d^2+1}{2}\right)^{\alpha_d}\left(\frac{\bar{b}^2+1}{2}\right)^{\alpha_{d+1}} \cdot \mathbb{E}_x[|x_1^{\alpha_1}\ldots x_d^{\alpha_d}\chi_T(x)|] \\
&= \varepsilon^{k/2}\text{poly}_T(\bar{w}, \bar{b})
\end{aligned}$$

$\square$

Lemma B.5. *For any $T \in \mathbb{N}^d$ such that $|T| \leqslant p$ we have*

$$\widehat{\phi}_{w,b}(T) = \varepsilon^{|T|/2} \cdot \frac{\sigma^{(|T|)}(0)}{|T|!} \binom{|T|}{t_1 \dots t_d} \bar{w}_1^{t_1} \dots \bar{w}_d^{t_d} + O(\varepsilon^{(|T|+1)/2} \cdot poly_T(\bar{w}, \bar{b}))$$

Here, $\bar{w} = \varepsilon^{-1/2} w$, $\bar{b} = \varepsilon^{-1/2} b$, so that the distribution of $\bar{w}$ and $\bar{b}$ does not depend on ε: $\bar{w} \sim \mathcal{N}(0, I_d)$, $\bar{b} \sim \mathcal{N}(0,1)$.

Proof. Using Taylor series, we get

$$\phi_{w,b}(x) = \sigma(\langle w, x\rangle + b) = \sigma(G) = \sigma(0) + \sigma'(0) G + \ldots + \frac{\sigma^{(|T|)}(0)}{|T|!} G^{|T|} + \frac{\sigma^{(|T|+1)}(\xi)}{(|T|+1)!} G^{|T|+1}$$

where $|\xi| \leqslant |G|$. Note that the expansion $\phi_{w,b}(x) = \sum_T \widehat{\phi}_{w,b}(T) \chi_T(x)$ is the basis of Hermite polymonials $\chi_T(x)$ does not depend on the distribution on the input space $\mathbb{R}^d$. Hence, for simplicity, we can assume the Gaussian distribution: $x \sim \mathcal{N}(0, I_d)$, which allows us to express the Hermite coefficients using the dot product: $\widehat{\phi}_{w,b}(T) = \mathbb{E}_x[\phi_{w,b}(x)\chi(T)]$. Thus, by Lemma B.3 we obtain

$$\widehat{\phi}_{w,b}(T) = \mathbb{E}_x[\frac{\sigma^{(|T|)}(0)}{|T|!} G^{|T|} \chi_T(x) + \frac{\sigma^{(|T|+1)}(\xi)}{(|T|+1)!} G^{|T|+1} \chi_T(x)] = \mathbb{E}_x[A] + \mathbb{E}_x[B]$$

where $A = \frac{\sigma^{(|T|)}(0)}{|T|!} G^{|T|} \chi_T(x)$ and $B = \frac{\sigma^{(|T|+1)}(\xi)}{(|T|+1)!} G^{|T|+1} \chi_T(x)$. For the first term we get

$$\begin{aligned}\mathbb{E}_x[A] = \frac{\sigma^{(|T|)}(0)}{|T|!} \mathbb{E}_x[(\langle w, x\rangle + b)^{|T|} \chi_T(x)] &= \varepsilon^{|T|/2} \frac{\sigma^{(|T|)}(0)}{|T|!} \mathbb{E}_x[(\langle \bar{w}, x\rangle + \bar{b})^{|T|} \chi_T(x)] \\ &= \varepsilon^{|T|/2} \frac{\sigma^{(|T|)}(0)}{|T|!} \binom{|T|}{t_1 \dots t_d} \bar{w}_1^{t_1} \dots \bar{w}_d^{t_d}\end{aligned}$$

For the second term, we can estimate

$$|\mathbb{E}_x[B]| \leqslant \mathbb{E}_x[|B|] = \frac{1}{(|T|+1)!} \mathbb{E}_x[|\sigma^{(|T|+1)}(\xi) G^{|T|+1} \chi_T(x)|]$$

By assumption 4.1 we have

$$|\sigma^{(|T|+1)}(\xi)| \leqslant C(\xi^l + 1) \leqslant C(|\xi|^l + 1) \leqslant C(|G|^l + 1)$$

Substituting, we obtain:

$$|\mathbb{E}_x[B]| \leqslant \frac{C}{(|T|+1)!} \left(\mathbb{E}_x[|G^{|T|+1} \chi_T(x)|] + \mathbb{E}_x[|G^{|T|+l+1} \chi_T(x)|] \right)$$

Applying Lemma B.4 to the expectations above, we proceed

$$\begin{aligned}|\mathbb{E}_x[B]| &\leqslant \frac{C}{(|T|+1)!} \left(\varepsilon^{(|T|+1)/2} \text{poly}_T^{(1)}(\bar{w}, \bar{b}) + \varepsilon^{(|T|+l+1)/2} \text{poly}_T^{(2)}(\bar{w}, \bar{b}) \right) \\ &\leqslant \frac{C}{(|T|+1)!} \cdot \varepsilon^{(|T|+1)/2} \left(\text{poly}_T^{(1)}(\bar{w}, \bar{b}) + \text{poly}_T^{(2)}(\bar{w}, \bar{b}) \right) \\ &= \varepsilon^{(|T|+1)/2} \text{poly}_T(\bar{w}, \bar{b})\end{aligned}$$

which completes the proof. □

Lemma B.6. *If $g \in \Pi_p(\mathbb{R}^d)$, then*

$$\widehat{g}_{\leqslant p}^\top \Phi^{-1} \widehat{g}_{\leqslant p} = \Theta(\varepsilon^{-def(g)})$$

Proof. By the previous lemma we know that

$$\Phi = \text{Gram}\left\{\widehat{\phi}_{w,b}(T) = \varepsilon^{|T|/2} c_T \bar{w}_1^{t_1} \dots \bar{w}_d^{t_d} + O(\varepsilon^{(|T|+1)/2} \cdot \text{poly}_T(\bar{w}, \bar{b})),\ T \in \mathbb{N}^d_{\leqslant p}\right\}$$

where $c_T = \frac{\sigma^{(|T|)}(0)}{|T|!}\binom{|T|}{t_1 \dots t_d} \neq 0$ given that $\sigma^{(|T|)}(0) \neq 0$ by Assumption 4.1. Define

$$A = \text{Gram}\left\{\varepsilon^{-|T|/2}\widehat{\phi}_{w,b}(T) = c_T \bar{w}_1^{t_1} \dots \bar{w}_d^{t_d} + O(\varepsilon^{1/2} \cdot \text{poly}_T(\bar{w}, \bar{b})),\ T \in \mathbb{N}^d_{\leqslant p}\right\}$$

Then A and Φ are connected by

$$\Phi_{i,j} = \varepsilon^{(|T_i|+|T_j|)/2} \cdot A_{i,j} \tag{5}$$

$$\Phi^{-1}_{i,j} = \varepsilon^{-(|T_i|+|T_j|)/2} \cdot A^{-1}_{i,j} \tag{6}$$

(the second can be established, for example, by Cramer's rule). Next, define

$$\tilde{A} = \text{Gram}\left\{c_T \bar{w}_1^{t_1} \dots \bar{w}_d^{t_d},\ T \in \mathbb{N}^d_{\leqslant p}\right\}$$

e.g. we dropped the reminders from the Gram basis elements of A. Then we have that $A_{i,j} = \tilde{A}_{i,j} + O(\varepsilon^{1/2})\ \forall i, j$ (here, we use that the expectation of any polynomial in $\bar{w}, \bar{b}$ is finite). It gives us $\det(A) = \det(\tilde{A}) + O(\varepsilon^{1/2})$. Besides, denoting by C and $\tilde{C}$ the cofactor matrices of A and $\tilde{A}$ respectively, we also have $C_{i,j} = \tilde{C}_{i,j} + O(\varepsilon^{1/2})\ \forall i, j$. Finally, $\det(\tilde{A}) \neq 0$ since the basis elements of $\tilde{A}$ are linearly independent. Combining all together, we have

$$A^{-1}_{i,j} = \frac{C_{j,i}}{\det(A)} = \frac{\tilde{C}_{j,i} + O(\varepsilon^{1/2})}{\det(\tilde{A}) + O(\varepsilon^{1/2})} = \tilde{A}^{-1}_{i,j} + O(\varepsilon^{1/2})$$

Combining with (6), we get

$$\Phi^{-1}_{i,j} = \varepsilon^{-(|T_i|+|T_j|)/2} \cdot A^{-1}_{i,j} = \varepsilon^{-(|T_i|+|T_j|)/2} \cdot \tilde{A}^{-1}_{i,j} + O(\varepsilon^{-(|T_i|+|T_j|-1)/2}) \tag{7}$$

As a corollary, we may estimate

$$\Phi^{-1}_{i,j} = O(\varepsilon^{-(|T_i|+|T_j|)/2}) \tag{8}$$

Now consider any fixed $g \in \Pi_p(\mathbb{R}^d)$. Denote $s = \deg(g)$, then $\widehat{g}(T) = 0\ \forall T : |T| > s$. Hence,

$$\widehat{g}_{\leqslant p}^\top \Phi^{-1} \widehat{g}_{\leqslant p} = \sum_{|T|,|T'| \leqslant s} \widehat{g}(T)\widehat{g}(T')\Phi^{-1}_{T,T'}$$

Note that if $|T| < s$ or $|T'| < s$ then $(|T| + |T'|)/2 \leqslant s - 1/2$ and from (8) we get $\Phi^{-1}_{T,T'} = O(\varepsilon^{-s+1/2})$. Thus, we can estimate

$$\widehat{g}_{\leqslant p}^\top \Phi^{-1} \widehat{g}_{\leqslant p} = \sum_{|T|,|T'|=s} \widehat{g}(T)\widehat{g}(T')\Phi^{-1}_{T,T'} + O(\varepsilon^{-s+1/2}) \overset{(7)}{=} \varepsilon^{-s} \sum_{|T|,|T'|=s} \widehat{g}(T)\widehat{g}(T')\tilde{A}^{-1}_{T,T'} + O(\varepsilon^{-s+1/2}) \tag{9}$$

Now define $g' \in \Pi_p(\mathbb{R}^d)$ by setting $\widehat{g}'(T) = \widehat{g}(T)$ if $|T| = s$ and $\widehat{g}'(T) = 0$ otherwise. Then

$$\sum_{|T|,|T'|=s} \widehat{g}(T)\widehat{g}(T')\tilde{A}^{-1}_{T,T'} = (\widehat{g}'_{\leqslant p})^\top \tilde{A}^{-1} \widehat{g}'_{\leqslant p} \tag{10}$$

Note that $\tilde{A} \succ 0$ since A is a Gram matrix of linearly independent set of functions. Thus, $\tilde{A}^{-1} \succ 0$. Moreover, since $deg(g) = s$, we have $\widehat{g}'_{\leqslant p} \neq 0$. Combining these, we conclude that the value of (10) is strictly positive. Denoting this value by $C_1 > 0$ and substituting it into (9), we obtain

$$\widehat{g}_{\leqslant p}^\top \Phi^{-1} \widehat{g}_{\leqslant p} = C_1 \varepsilon^{-s} + O(\varepsilon^{-s+1/2}) = \Theta(\varepsilon^{-s})$$

which completes the proof. ∎

Corollary B.7. *There exist $c, \varepsilon_0 > 0$ such that $\forall \varepsilon < \varepsilon_0$ we have:*

$$\Phi^{-1} \succeq cD_\varepsilon$$

where $D_\varepsilon = \operatorname{diag}(\{\varepsilon^{-|T_i|}, T_i \in \mathbb{N}^d_{\leqslant p}\})$

Proof. We can write (6) in matrix form as

$$\Phi^{-1} = D_\varepsilon^{1/2} A^{-1} D_\varepsilon^{1/2} \tag{11}$$

Since $\tilde{A}^{-1} \succ 0$, it holds that $\tilde{A}^{-1} \succ cI$ where $c = \lambda_{min}(\tilde{A}^{-1})/2$. Combining with $A^{-1} \to \tilde{A}^{-1}$, we obtain that $A^{-1} \succ cI$ for small enough ε. Substituting in (11), we proceed

$$\Phi^{-1} = D_\varepsilon^{1/2} A^{-1} D_\varepsilon^{1/2} \succeq D_\varepsilon^{1/2} (cI) D_\varepsilon^{1/2} = cD_\varepsilon$$

which completes the proof. □

Proof of Theorem 4.2. The first statement of the theorem is given by Lemma B.1. We now turn to the proof of the second statement. Define the set of polynomial interpolators $\mathcal{F}_{\text{int}}$ as

$$\mathcal{F}_{\text{int}} = \left\{ h \in \Pi_p(\mathbb{R}^d) : \forall x \in \mathcal{U}^c,\, h(x) = f(x) \right\} .$$

Since $f \in \Pi_p(\mathbb{R}^d)$, we have $\mathcal{F}_{\text{int}} \neq \varnothing$. Define the matrix $F \in \mathbb{R}^{|\mathbb{N}^d_{\leqslant p}| \times N}$ by setting $F_{ij} = \frac{1}{\sqrt{N}} \widehat{\phi}_{w_j, b_j}(T_i)$, where $j \in \{1, \ldots, N\}$ and $T_1, \ldots, T_{|\mathbb{N}^d_{\leqslant p}|}$ are enumerated elements of $\mathbb{N}^d_{\leqslant p}$. Then the Hermite coefficients of the random features model can be expressed as $\widehat{f}_{\text{RF}}(a) = Fa$.

Consider N large enough so that any interpolator $g \in \mathcal{F}_{\text{int}}$ can be expressed by the random features model (such N exists w.h.p. by Lemma B.1). Then the equation

$$Fa = \widehat{g} \tag{12}$$

has solution in a for any $g \in \mathcal{F}_{\text{int}}$. Moreover, provided that the matrix $FF^\top \in \mathbb{R}^{|\mathbb{N}^d_{\leqslant p}| \times |\mathbb{N}^d_{\leqslant p}|}$ is invertible, the minimum-norm solution $a(g)$ of (12) is given by

$$a(g) = F^\dagger \widehat{g} = F^\top (FF^\top)^{-1} \widehat{g} \tag{13}$$

and we get

$$\|a(g)\|^2 = \widehat{g}^\top (FF^\top)^{-1} \widehat{g} \,. \tag{14}$$

Let us show that $FF^\top$ is indeed invertible (w.h.p.). We have

$$\begin{aligned}
(FF^\top)_{i,j} &= \sum_{k=1}^N F_{i,k} F^\top_{k,j} = \sum_{k=1}^N F_{i,k} F_{j,k} \\
&= \frac{1}{N} \sum_{k=1}^N \widehat{\phi}_{w_k, b_k}(T_i) \widehat{\phi}_{w_k, b_k}(T_j) \\
&\xrightarrow[N \to \infty]{\text{a.s.}} \mathbb{E}_{w,b}[\widehat{\phi}_{w,b}(T_i) \widehat{\phi}_{w,b}(T_j)]
\end{aligned}$$

where the last step follows from the Strong Law of Large Numbers (SLLN). To be able to use the SLLN, we have to check that $\mathbb{E}_{w,b}[|\widehat{\phi}_{w,b}(T_i) \widehat{\phi}_{w,b}(T_j)|] < \infty$. For this, we use the Cauchy-Schwarz inequality:

$$\begin{aligned}
&\mathbb{E}_{w,b}[|\widehat{\phi}_{w,b}(T_i) \widehat{\phi}_{w,b}(T_j)|] && (15) \\
&\leqslant \sqrt{\mathbb{E}_{w,b}[\widehat{\phi}_{w,b}(T_i)^2]} \cdot \sqrt{\mathbb{E}_{w,b}[\widehat{\phi}_{w,b}(T_j)^2]} \,. && (16)
\end{aligned}$$

The finiteness of the right-hand side follows (at least for small enough ε) from Lemma B.5 after noting that the expectation of any polynomial in $\bar{w}, \bar{b}$ is finite, where $\bar{w} \sim \mathcal{N}(0, I_d)$, $\bar{b} \sim \mathcal{N}(0,1)$. In the following, we consider ε small enough for (15)–(16) to hold for any i, j.

Thus, we get that $FF^\top \overset{a.s.}{\to} \Phi$, where $\Phi \in \mathbb{R}^{|\mathbb{N}^d_{\leqslant p}| \times |\mathbb{N}^d_{\leqslant p}|}$ is a deterministic matrix defined by

$$\Phi_{ij} = \mathbb{E}_{w,b}[\widehat{\phi}_{w,b}(T_i)\widehat{\phi}_{w,b}(T_j)] \tag{17}$$

Let us show that the matrix Φ is invertible. Note that Φ is the Gram matrix for the set of functions $\{(w,b) \mapsto \widehat{\phi}_{w,b}(T),\ T \in \mathbb{N}^d_{\leqslant p}\}$ in $L^2(\mathbb{R}^{d+1}, \gamma_{d+1})$ space. Hence, if matrix Φ were degenerate, it would mean that the functions $\{(w,b) \mapsto \widehat{\phi}_{w,b}(T),\ T \in \mathbb{N}^d_{\leqslant p}\}$ are linearly dependent. Denote $k = |\mathbb{N}^d_{\leqslant p}|$. Then there exists a linear subspace L of dimension $\dim(L) \leqslant k-1$ such that for all w, b we have $\widehat{\phi}_{w,b} \in L$. This implies that $\widehat{f}_{\text{RF}} \in L$ for all N, $\{w_i\}_{i=1}^N$, $\{b_i\}_{i=1}^N$, $\{a_i\}_{i=1}^N$. Therefore, we have $\dim(\text{im}(\widehat{f}_{\text{RF}})) \leqslant k-1$ thus $\dim(\text{im}(f_{\text{RF}})) = \dim(\text{im}(\widehat{f}_{\text{RF}})) \leqslant k-1 < k = \dim(\Pi_p(\mathbb{R}^d))$. The last inequality shows that $\text{im}(f_{\text{RF}}) \neq \Pi_p(\mathbb{R}^d)$, and this statement holds for all N, $\{w_i\}_{i=1}^N$, $\{b_i\}_{i=1}^N$. Thus, we get a contradiction with Lemma B.1. This proves that matrix Φ must be invertible.

Since Φ is invertible and $FF^\top \to \Phi$ as $N \to \infty$, the matrix $FF^\top$ must be invertible for large enough N and $(FF^\top)^{-1} \to \Phi^{-1}$. Thus, we justified (13)–(14) and from (14) can deduce

$$\|a(g)\|^2 \xrightarrow[N\to\infty]{a.s.} \widehat{g}^\top \Phi^{-1} \widehat{g}\,. \tag{18}$$

Recall that a^* denotes the minimum norm interpolating solution. Thus, for finite N, $f_{\text{RF}}(a^*)$ is the minimizer of (14) over $g \in \mathcal{F}_{\text{int}}$. Besides, denote the minimizer of (18) over $g \in \mathcal{F}_{\text{int}}$ by g_ε. Since $(FF^\top)^{-1} \succ 0$ (for large enough N), $\Phi^{-1} \succ 0$, $(FF^\top)^{-1} \to \Phi^{-1}$ as $N \to \infty$, and since $\mathcal{F}_{\text{int}}$ is an affine subspace, by Lemma B.2 we get that $\widehat{f}_{\text{RF}}(a^*) \to \widehat{g}_\varepsilon$ as $N \to \infty$ for any small enough $\varepsilon > 0$, which implies $f_{\text{RF}}(a^*) \to g_\varepsilon$.

It remains to show that $\text{dist}(g_\varepsilon, \Pi_{p_*}) \to 0$ as $\varepsilon \to 0$. Consider $h \in \mathcal{F}_{\text{int}}$ such that $\deg(h) = p_*$. Then by Lemma B.6 we have $\widehat{h}^\top \Phi^{-1} \widehat{h} = \Theta(\varepsilon^{-p_*})$, which implies $\exists c_1 > 0$: $\widehat{h}^\top \Phi^{-1}\widehat{h} \leqslant c_1 \varepsilon^{-p_*}$ for any small enough $\varepsilon > 0$. Since g_ε is the minimizer of (18), we can estimate

$$\widehat{g}_\varepsilon^\top \Phi^{-1} \widehat{g}_\varepsilon \leqslant \widehat{h}^\top \Phi^{-1} \widehat{h} \leqslant c_1 \varepsilon^{-p_*} \tag{19}$$

On the other hand, from Corollary B.7, there exists $c_2 > 0$ such that

$$\begin{aligned}\widehat{g}_\varepsilon^\top \Phi^{-1} \widehat{g}_\varepsilon &\geqslant c_2 \sum_{|T| \leqslant p} \widehat{g}_\varepsilon(T)^2 \varepsilon^{-|T|} \\ &\geqslant c_2 \sum_{|T|=k} \widehat{g}_\varepsilon(T)^2 \varepsilon^{-k} = c_2 e_\varepsilon(k) \varepsilon^{-k}\end{aligned}$$

where we define $e_\varepsilon(k) = \sum_{|T|=k} \widehat{g}_\varepsilon(T)^2$ - the energy of the degree-k monomials of g_ε. Combining this with (19), we obtain

$$c_2 e_\varepsilon(k) \varepsilon^{-k} \leqslant \widehat{g}_\varepsilon^\top \Phi^{-1} \widehat{g}_\varepsilon \leqslant c_1 \varepsilon^{-p_*} \Rightarrow e_\varepsilon(k) \leqslant \frac{c_1}{c_2} \varepsilon^{k-p_*}\,.$$

For $k > p_*$ we have $\varepsilon^{k-p_*} \to 0$, and thus $e_\varepsilon(k) \to 0$. As $\text{dist}(g_\varepsilon, \Pi_{p_*}(\mathbb{R}^d))$ is bounded by $\sum_{k>p_*} e_\varepsilon(k)$ up to a constant, this concludes the proof. ☐

C. Proof of Example 5.3

Lemma C.1. *Let $T = (t_1, \ldots, t_d), T' = (t'_1, \ldots, t'_d) \in \mathbb{N}^d$ are such that $\exists i \in [d]$: $t_i \not\equiv t'_i$ (mod 2). Then it holds*

$$\mathbb{E}_{w,b}[\widehat{\phi}_{w,b}(T)\widehat{\phi}_{w,b}(T')] = 0$$

Proof. Denote by w_{-i} and x_{-i} the vectors w and x respectively with flipped i-th coordinate. Note that

$$\phi_{w_{-i},b}(x) = \sigma(\langle w_{-i}, x\rangle + b) = \sigma(\langle w, x_{-i}\rangle + b) = \phi_{w,b}(x_{-i})$$

Suppose that $\phi_{w,b}(x)$ has the following Hermite decomposition:

$$\phi_{w,b}(x) = \sum_{T \in \mathbb{N}^d} \widehat{\phi}_{w,b}(T)\chi_T(x)$$

where $\chi_T(x) = \prod_{i=1}^d H_{t_i}(x_i)$. Note that the Hermite polynomial $H_t(x_t)$ is an odd function for odd t and even function for even t. Thus, we have

$$\chi_T(x_{-i}) = \begin{cases} \chi_T(x), & t_i \equiv 0 \ (mod\ 2) \\ -\chi_T(x), & t_i \equiv 1 \ (mod\ 2) \end{cases}$$

Thus, for the function $\phi_{w_{-i},b}(x)$, we get:

$$\phi_{w_{-i},b}(x) = \phi_{w,b}(x_{-i}) = \sum_{T \in \mathbb{N}^d, t_i \equiv 0\ (mod\ 2)} \widehat{\phi}_{w,b}(T)\chi_T(x) - \sum_{T \in \mathbb{N}^d, t_i \equiv 1\ (mod\ 2)} \widehat{\phi}_{w,b}(T)\chi_T(x)$$

which shows that

$$\widehat{\phi}_{w_{-i},b}(T) = \begin{cases} \widehat{\phi}_{w,b}(T), & t_i \equiv 0 \ (mod\ 2) \\ -\widehat{\phi}_{w,b}(T), & t_i \not\equiv 0 \ (mod\ 2) \end{cases} \tag{20}$$

Finally, consider $i \in [d]$ for which $t_i \not\equiv t'_i \ (mod\ 2)$. Then

$$\mathbb{E}_{w,b}[\widehat{\phi}_{w,b}(T)\widehat{\phi}_{w,b}(T')] = \mathbb{E}_{w,b}[\widehat{\phi}_{w_{-i},b}(T)\widehat{\phi}_{w_{-i},b}(T')] = -\mathbb{E}_{w,b}[\widehat{\phi}_{w,b}(T)\widehat{\phi}_{w,b}(T')]$$

Here, the first equality comes from the fact that w and w_{-i} have the same distribution, and the second equality comes from (20). Hence, we obtained

$$\mathbb{E}_{w,b}[\widehat{\phi}_{w,b}(T)\widehat{\phi}_{w,b}(T')] = -\mathbb{E}_{w,b}[\widehat{\phi}_{w,b}(T)\widehat{\phi}_{w,b}(T')]$$

which completes the proof. □

Proposition C.2. *Let the random features model be trained in sparse setting and diverging d regime with $\sigma(x) = (1+x)^2$ activation. Then it converges to the minimizer of:*

$$\sum_{i=1}^d \widehat{g}(2e_i)^2 \cdot \frac{d^2}{4} + \sum_{i<j} \widehat{g}(e_i+e_j)^2 \cdot \frac{d^2}{4} \tag{21}$$

$$+ \sum_{i=1}^d \widehat{g}(e_i)^2 \cdot \frac{d}{4} + \widehat{g}(0)^2 \cdot \frac{d}{6} + \sum_{i<j} \widehat{g}(2e_i)\widehat{g}(2e_j) \cdot \frac{d}{6} + \sum_{i=1}^d \widehat{g}(2e_i)\widehat{g}(0) \cdot \left(-\frac{\sqrt{2}}{3}d\right) \tag{22}$$

over functions g that interpolate the training data.

Proof. The arguments in Theorem 4.2 showing that the random feature model converges to the minimizer of $\widehat{g}^\top \Phi^{-1} \widehat{g}$ in small feature regime (where matrix Φ is defined in (17)) transfer directly to the sparse regime (but now we will have this convergence for fixed d instead of fixed ε). Let us compute this quadratic form explicitly. We have

$$\begin{aligned} \phi_{w,b}(x) = \sigma(\langle w, x\rangle + b) &= \left(\sum_{i=1}^d w_i x_i + b + 1\right)^2 \\ &= \sum_{i=1}^d w_i^2 x_i^2 + 2\sum_{i<j} w_i w_j x_i x_j + (b+1)^2 + 2\sum_{i=1}^d w_i (b+1) x_i \\ &= \sum_{i=1}^d w_i^2 \sqrt{2} \frac{x_i^2 - 1}{\sqrt{2}} + 2\sum_{i<j} w_i w_j x_i x_j + 2\sum_{i=1}^d w_i (b+1) x_i + (b+1)^2 + \sum_{i=1}^d w_i^2 \end{aligned}$$

Thus the Hermite coefficients of the random feature $\phi_{w,b}$ are given by:

$$\widehat{\phi}(2,0,\dots,0) = w_1^2\sqrt{2} \Rightarrow \mathbb{E}[\widehat{\phi}(2,0,\dots,0)^2] = 2\mathbb{E}[w_1^4] = \frac{6}{d^2}$$
$$\widehat{\phi}(1,1,0,\dots,0) = 2w_1w_2 \Rightarrow \mathbb{E}[\widehat{\phi}(1,1,0,\dots,0)^2] = 4\mathbb{E}[w_1^2w_2^2] = \frac{4}{d^2}$$
$$\widehat{\phi}(1,0,\dots,0) = 2w_1(b+1) \Rightarrow \mathbb{E}[\widehat{\phi}(1,0,\dots,0)^2] = 4\mathbb{E}[w_1^2(b+1)^2] = 4\cdot\frac{1}{d}\cdot(1+\frac{1}{d}) = 4(\frac{1}{d}+\frac{1}{d^2})$$
$$\widehat{\phi}(0,0,\dots,0) = (b+1)^2 + \sum_{i=1}^{d} w_i^2 \Rightarrow \mathbb{E}[\widehat{\phi}(0,0,\dots,0)^2] \overset{(a)}{=} 4 + \frac{10}{d} + \frac{3}{d^2}$$

Here, (a) comes from

$$\mathbb{E}[\widehat{\phi}(0,0,\dots,0)^2] = \mathbb{E}\left[\left((b+1)^2 + \sum_{i=1}^{d} w_i^2\right)^2\right]$$
$$= \mathbb{E}[(b+1)^4 + \sum_{i=1}^{d} w_i^4 + 2\sum_{i<j} w_i^2w_j^2 + 2\sum_{i=1}^{d} w_i^2(b+1)^2] = (1+\frac{6}{d}+\frac{3}{d^2}) + d\cdot\frac{3}{d^2} + d(d-1)\frac{1}{d^2} + 2d\cdot\frac{1}{d}(1+\frac{1}{d})$$
$$= (1+\frac{6}{d}+\frac{3}{d^2}) + \frac{3}{d} + (1-\frac{1}{d}) + (2+\frac{2}{d})$$
$$= 4 + \frac{10}{d} + \frac{3}{d^2}$$

For the cross-terms, we have

$$\mathbb{E}[\widehat{\phi}(2,0,\dots,0)\widehat{\phi}(0,2,\dots,0)] = 2\mathbb{E}[w_1^2w_2^2] = \frac{2}{d}$$
$$\mathbb{E}[\widehat{\phi}(2,0,\dots,0)\widehat{\phi}(0,0,\dots,0)] = \sqrt{2}\mathbb{E}[w_1^2(b+1)^2 + w_1^4 + \sum_{i=2}^{d} w_1^2w_i^2] = \sqrt{2}\left(\frac{1}{d}(1+\frac{1}{d}) + \frac{3}{d^2} + (d-1)\cdot\frac{1}{d^2}\right)$$
$$= \sqrt{2}\left(\frac{2}{d}+\frac{3}{d^2}\right)$$

All other cross-terms equal to zero by Lemma C.1. Thus, matrix Φ is block-diagonal with the only non-unit block corresponding to the coefficients $(2,0,\dots,0),(0,2,\dots,0),(0,0,\dots,2),(0,0,\dots,0)$ ($d+1$ coefficient in the block in total). Thus, this $(d+1)\times(d+1)$ block takes the form:

$$\begin{bmatrix} \frac{6}{d^2} & \frac{2}{d^2} & \dots & \frac{2}{d^2} & \sqrt{2}\left(\frac{2}{d}+\frac{3}{d^2}\right) \\ \frac{2}{d^2} & \frac{6}{d^2} & \dots & \frac{2}{d^2} & \sqrt{2}\left(\frac{2}{d}+\frac{3}{d^2}\right) \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ \frac{2}{d^2} & \frac{2}{d^2} & \dots & \frac{6}{d^2} & \sqrt{2}\left(\frac{2}{d}+\frac{3}{d^2}\right) \\ \sqrt{2}\left(\frac{2}{d}+\frac{3}{d^2}\right) & \sqrt{2}\left(\frac{2}{d}+\frac{3}{d^2}\right) & \dots & \sqrt{2}\left(\frac{2}{d}+\frac{3}{d^2}\right) & \left(4+\frac{10}{d}+\frac{3}{d^2}\right) \end{bmatrix}$$

Exploiting the permutation symmetry of the first d Hermite coefficients in this matrix, we can search for its inverse in the following form:

$$\begin{bmatrix} x & y & \dots & y & z \\ y & x & \dots & y & z \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ y & y & \dots & x & z \\ z & z & \dots & z & t \end{bmatrix}$$

From the condition that the product of the formal matrix with the later one must be I_{d+1}, we obtain the following 4 linear equations in 4 unknown variables:

$$\frac{6}{d^2}x + 2\left(\frac{1}{d} - \frac{1}{d^2}\right)y + \sqrt{2}\left(\frac{2}{d} + \frac{3}{d^2}\right)z = 1$$
$$\frac{2}{d^2}x + 2\left(\frac{1}{d} + \frac{1}{d^2}\right)y + \sqrt{2}\left(\frac{2}{d} + \frac{3}{d^2}\right)z = 0$$
$$2\left(\frac{1}{d} + \frac{2}{d^2}\right)z + \sqrt{2}\left(\frac{2}{d} + \frac{3}{d^2}\right)t = 0$$
$$\sqrt{2}\left(2 + \frac{3}{d}\right)z + \left(4 + \frac{10}{d} + \frac{3}{d^2}\right)t = 1$$

Solving this system, we obtain

$$x = \frac{d^2}{4} + O(d)$$
$$y = \frac{d}{12} + O(1)$$
$$z = -\frac{\sqrt{2}d}{6} + O(1)$$
$$t = \frac{d}{6} + O(1)$$

Combining with

$$\left(\mathbb{E}[\widehat{\phi}(1,1,0,\dots,0)^2]\right)^{-1} = \frac{d^2}{4}$$
$$\left(\mathbb{E}[\widehat{\phi}(1,0,\dots,0)^2]\right)^{-1} = \frac{d}{4} + O(1)$$

we obtain:

$$\widehat{g}^\top \Phi^{-1}\widehat{g} \approx \sum_{i=1}^{d} \widehat{g}(2e_i)^2 \cdot \frac{d^2}{4} + \sum_{i<j} \widehat{g}(e_i+e_j)^2 \cdot \frac{d^2}{4}$$
$$+ \sum_{i=1}^{d} \widehat{g}(e_i)^2 \cdot \frac{d}{4} + \widehat{g}(0)^2 \cdot \frac{d}{6} + \sum_{i<j} \widehat{g}(2e_i)\widehat{g}(2e_j) \cdot \frac{d}{6} + \sum_{i=1}^{d} \widehat{g}(2e_i)\widehat{g}(0) \cdot \left(-\frac{\sqrt{2}}{3}d\right)$$

□

Proof of Example 5.3. Assume that g is an interpolator of the training data. We show that

$$g(x) = \widehat{g}(0) + \widehat{g}(e_1)x_1 + \widehat{g}(2e_1)\frac{x_1^2 - 1}{\sqrt{2}} + \sum_{i \geqslant 2} \left(\widehat{g}(e_i)x_i + \widehat{g}(e_1+e_i)x_1x_i\right) ,$$

with the constraints that $\widehat{g}(0) + \widehat{g}(e_1) = 1$ and $\widehat{g}(e_i) + \widehat{g}(e_1 + e_i) = 0$ for all $i \geqslant 2$.

Recall that the support of this distribution contains a subset of the form $\{1\} \times S_2 \times \cdots \times S_d$ where $S_2, \dots, S_d$ have cardinality at least 3. We apply Theorem 1.1 of (Alon, 1999). There exists multivariate polynomials $h_1, h_2, \dots, h_d$ such that $g = \sum_i h_i g_i$ with $g_1(x) = x_1 - 1$ and for $i \geqslant 2$, $g_i(x) = \Pi_{s \in S_i}(x - x_i)$. Moreover, the degree of h_1 is at most $\deg f - \deg g_1 = 1$ and the degree of h_i ($i \geqslant 2$) is at most $\deg f - \deg g_i = -1$. Thus $h_2 = \cdots = h_d = 0$. We thus get $g(x) = h_1(x)(x_1 - 1)$ with $h_1(x)$ affine, which is equivalent to the above statement.

To sum up, we minimize

$$\widehat{g}^\top \Phi^{-1}\widehat{g} \approx \widehat{g}(2e_1)^2 \cdot \frac{d^2}{4} + \sum_{i \geqslant 2} \widehat{g}(e_1+e_i)^2 \cdot \frac{d^2}{4} + \widehat{g}(e_1)^2 \cdot \frac{d}{4}$$
$$+ \sum_{i \geqslant 2} \widehat{g}(e_i)^2 \cdot \frac{d}{4} + \widehat{g}(0)^2 \cdot \frac{d}{6} + \widehat{g}(2e_1)\widehat{g}(0) \cdot \left(-\frac{\sqrt{2}}{3}d\right)$$

in the variables $\widehat{g}(0), \widehat{g}(e_1), \widehat{g}(2e_1), \widehat{g}(e_1+e_i), \widehat{g}(e_i)$, $i \geqslant 2$ with constraints $\widehat{g}(0)+\widehat{g}(e_1)=1$ and $\widehat{g}(e_i)+\widehat{g}(e_1+e_i)=0$.

This optimization problem is separable in the groups of variables $\{\widehat{g}(0), \widehat{g}(e_1), \widehat{g}(2e_1)\}$ and $\{\widehat{g}(e_1+e_i), \widehat{g}(e_i)\}$ (both the constraints and the objective are separable). The second optimization problem is obvious and leads to the unique solution $\widehat{g}(e_i)=\widehat{g}(e_1+e_i)=0$, $i \geqslant 2$. Simplifying the scaling of the objective, we are left with the optimization problem of minimizing

$$\widehat{g}(2e_1)^2 \cdot \frac{d}{4} + \widehat{g}(e_1)^2 \cdot \frac{1}{4} + \widehat{g}(0)^2 \cdot \frac{1}{6} + \widehat{g}(2e_1)\widehat{g}(0) \cdot \left(-\frac{\sqrt{2}}{3}\right)$$

under the constraint $\widehat{g}(0)+\widehat{g}(e_1)=1$.

Minimizing marginally in $\widehat{g}(2e_1)$, we obtain that $\widehat{g}(2e_1)=\widehat{g}(0)\frac{2\sqrt{2}}{3d}$. Substituting in the expression above, we minimize

$$\widehat{g}(e_1)^2 \cdot \frac{1}{4} + \widehat{g}(0)^2 \cdot \frac{1}{6} - \widehat{g}(0)^2 \frac{4}{9d}$$

under the constraint $\widehat{g}(0)+\widehat{g}(e_1)=1$. The last term has a negligible effect as $d \to \infty$ and thus the solution converges to the solution with $\widehat{g}(0)=\frac{3}{5}$, $\widehat{g}(e_1)=\frac{2}{5}$. Thus, the random feature model learns the function $f_{\mathrm{RF}}(x)=\frac{2}{5}x_1+\frac{3}{5}$. $\square$

D. Proof of Theorem 6.1

The proof follows a structure similar to the one of (Abbe et al., 2023) (and to the one of Section 4 and Appendix B): the strategy is to study the covariance matrix of the random features. The minimum degree bias follows from different scales (in d) of different polynomial components of the random features. Here we only outline the main differences with the previous proofs.

For functions $h : \mathbb{U}_n^d \to \mathbb{C}$, the appropriate decomposition is given by its discrete Fourier transform. It corresponds to the linear decomposition on the basis of monomials

$$\chi_{j_1,\dots,j_d}(x) = x_1^{j_1} \cdots x_d^{j_d} \, .$$

This basis is orthonormal in the Hermitian space $L^2(\mathbb{U}_n^d, \mathrm{Unif}(\mathbb{U}_n^d))$. More concretely, the discrete Fourier transform of $h : \mathbb{U}_n^d \to \mathbb{C}$ is $\widehat{h} : \{0,\dots,n-1\}^d \to \mathbb{C}$, such that

$$\begin{gathered}\widehat{h}(j_1,\dots,j_d) = \mathbb{E}_{x\sim\mathrm{Unif}(\mathbb{U}_n^d)}\left[h(x)\overline{x}_1^{j_1}\cdots\overline{x}_d^{j_d}\right] \, , \\ j_1,\dots,j_d \in \{0,\dots,n-1\} \, .\end{gathered}$$

The inverse Fourier transform states that

$$h(x) = \sum_{j_1,\dots,j_d\in\{0,\dots,n-1\}} \widehat{h}(j_1,\dots,j_d) x_1^{j_1}\cdots x_d^{j_d} \, .$$

We consider the discrete Fourier transform of the random feature $\phi_{w,b}(x) = \sigma\left(\langle w, x\rangle + b\right)$:

$$\widehat{\phi}_{w,b}(j) = \mathbb{E}_x\left[\phi_{w,b}(x)\overline{x}_1^{j_1}\cdots\overline{x}_d^{j_d}\right] \, .$$

Theorem 6.1 stems from the following proposition.

Proposition D.1. *Consider $j, j' \in \{0,\dots,n-1\}^d$, $j \neq j'$. Then*

1. $\mathbb{E}_{w,b}\left[\widehat{\phi}_{w,b}(j)\overline{\widehat{\phi}_{w,b}(j')}\right] = 0$*, and*
2. $\mathbb{E}_{w,b}\left[\left|\widehat{\phi}_{w,b}(j)\right|^2\right] = \Theta\left(d^{-|j|}\right)$ *as* $d \to \infty$.

The two points of this proposition correpond respectively to the points A4 and A3 in Lemma A.1 of (Abbe et al., 2023).

Proof. 1. From $j \neq j'$, we know that there exists $l \in \{1, \dots, d\}$ such that $j_l \neq j'_l$. Let R denote the rotation of the l-th root of unity

$$T : (x_1, \dots, x_d) \in \mathbb{U}_n^d \mapsto \\ (x_1, \dots, x_{l-1}, e^{i\frac{2\pi}{l}} x_l, x_{l+1}, \dots, x_n)\,.$$

We compute the effect of a rotation of w on the discrete Fourier transform of a random feature:

$$\widehat{\phi}_{Tw,b}(j) = \mathbb{E}_x \left[\phi_{Tw,b}(x) \overline{x}_1^{j_1} \cdots \overline{x}_d^{j_d}\right]\,.$$

Here we note that the uniform distribution of $\mathbb{U}_n^d$ is invariant under the map T, thus

$$\widehat{\phi}_{Tw,b}(j) = \mathbb{E}_x \left[\phi_{Tw,b}(Tx) \overline{(Tx)}_1^{j_1} \cdots \overline{(Tx)}_d^{j_d}\right]\,.$$

Moreover,

$$\begin{aligned}\langle Tw, Tx\rangle = \overline{w}_1 x_1 + \cdots + \overline{w}_{l-1} x_{l-1} \overline{e^{i\frac{2\pi}{n}} w_l} e^{i\frac{2\pi}{n}} x_l \\ + \overline{w}_{l+1} x_{l+1} + \cdots + \overline{w}_d w_d = \langle w, x\rangle\end{aligned}$$

and thus $\phi_{Tw,b}(Tx) = \phi_{w,b}(x)$. As a consequence, we have

$$\begin{aligned}\widehat{\phi}_{Tw,b}(j) &= \mathbb{E}_x \left[\phi_{w,b}(x) \overline{(Tx)}_1^{j_1} \cdots \overline{(Tx)}_d^{j_d}\right] \\ &= e^{-i\frac{2\pi}{n}} \mathbb{E}_x \left[\phi_{w,b}(x) \overline{x}_1^{j_1} \cdots \overline{x}_d^{j_d}\right] \\ &= e^{-i\frac{2\pi j_l}{n}} \widehat{\phi}_{w,b}(j)\,.\end{aligned}$$

We are ready to conclude. The distribution of w is invariant under the map T, thus

$$\begin{aligned}\mathbb{E}_{w,b}\left[\widehat{\phi}_{w,b}(j)\overline{\widehat{\phi}_{w,b}(j')}\right] &= \mathbb{E}_{w,b}\left[\widehat{\phi}_{Tw,b}(j)\overline{\widehat{\phi}_{Tw,b}(j')}\right] \\ &= e^{i\frac{2\pi(j'_l - j_l)}{n}} \mathbb{E}_{w,b}\left[\widehat{\phi}_{w,b}(j)\overline{\widehat{\phi}_{w,b}(j')}\right]\,.\end{aligned}$$

As $j_l \neq j'_l$ and $j_l, j'_l \in \{0, \dots, n-1\}$, we have $e^{i\frac{2\pi(j'_l - j_l)}{n}} \neq 1$. Thus it must be that $\mathbb{E}_{w,b}\left[\widehat{\phi}_{w,b}(j)\overline{\widehat{\phi}_{w,b}(j')}\right] = 0$.

2. We make a Taylor expansion of σ at 0:

$$\widehat{\phi}_{w,b}(j) = \mathbb{E}_x\left[\sigma(\langle w, x\rangle + b)\overline{\chi_j(x)}\right] = \sum_{k=0}^{\infty} \frac{\sigma^{(k)}(0)}{k!} \mathbb{E}_x\left[(\langle w, x\rangle + b)^k \overline{\chi_j(x)}\right]\,.$$

We make three cases depending on the index k of the sum:

- If $k < |j|$, then $(\langle w, x\rangle + b)^k$ is a polynomial of degree $< k \leqslant |j| = \deg \chi_j$ thus by orthogonality $\mathbb{E}_x\left[(\langle w, x\rangle + b)^k \overline{\chi_j(x)}\right] = 0$.
- If $k = |j|$, then

$$\begin{aligned}\mathbb{E}_x\left[(\langle w, x\rangle + b)^k \overline{\chi_j(x)}\right] &= \mathbb{E}_x\left[(\overline{w}_1 x_1 + \cdots + \overline{w}_d x_d + b)^k \overline{\chi_j(x)}\right] \\ &= \sum_{l_1 + \cdots + l_d + l_{d+1} = k} \binom{k}{l_1, \dots, l_d, l_{d+1}} \mathbb{E}_x\left[(\overline{w}_1 x_1)^{l_1} \cdots (\overline{w}_d x_d)^{l_d} b^{l_{d+1}} \overline{\chi_j(x)}\right] \\ &= \sum_{l_1 + \cdots + l_d + l_{d+1} = k} \binom{k}{l_1, \dots, l_d, l_{d+1}} \overline{w}_1^{l_1} \cdots \overline{w}_d^{l_d} b^{l_{d+1}} \mathbb{E}_x\left[x_1^{l_1 - j_1} \cdots x_d^{l_d - j_d}\right]\,.\end{aligned}$$

Note that $\mathbb{E}_x\left[x_1^{l_1-j_1}\cdots x_d^{l_d-j_d}\right]$ equals 1 if $l_1 \equiv j_1 \mod n, \ldots, l_d \equiv j_d \mod n$ and 0 otherwise. As $l_1+\cdots+l_d = k - l_{d+1} \leqslant k = j_1+\cdots+j_d$, this is possible if and only if $l_1 = j_1, \ldots, l_d = j_d$, $l_{d+1} = 0$. Thus

$$\mathbb{E}_x\left[(\langle w,x\rangle + b)^k\overline{\chi_j(x)}\right] = \binom{k}{j_1,\ldots,j_d}\overline{w}_1^{j_1}\cdots\overline{w}_d^{j_d} = \frac{1}{d^{|j|/2}}\binom{k}{j_1,\ldots,j_d}\overline{u}_1^{j_1}\cdots\overline{u}_d^{j_d}$$

where $u := d^{1/2}w$ (and thus $u_1,\ldots,u_d$ are i.i.d. with standard Gaussian distribution in the complex plane).

- If $k > |j|$, then

$$\mathbb{E}_x\left[(\langle w,x\rangle + b)^k\overline{\chi_j(x)}\right] = \frac{1}{d^{k/2}}\mathbb{E}_x\left[(\langle u,x\rangle + b)^k\overline{\chi_j(x)}\right] ,$$

where again $u = d^{1/2}w$ and $c := d^{1/2}b$ (and thus b has standard Gaussian distribution in the complex plane).

Putting these three points together, we obtain

$$\widehat{\phi}_{w,b}(j) = \frac{1}{d^{|j|/2}}\binom{k}{j_1,\ldots,j_d}\overline{u}_1^{j_1}\cdots\overline{u}_d^{j_d} + o\left(\frac{1}{d^{|j|/2}}\right) .$$

Thus

$$\mathbb{E}_{w,b}\left[\left|\widehat{\phi}_{w,b}(j)\right|^2\right] = \frac{1}{d^{|j|}}\binom{k}{j_1,\ldots,j_d}^2\mathbb{E}_w\left[|u_1|^{2j_1}\cdots|u_d|^{2j_d}\right] + o\left(\frac{1}{d^{|j|}}\right) = \Theta\left(\frac{1}{d^{|j|}}\right) .$$

$\square$