SYMMETRIC SINGLE INDEX LEARNING

Aaron Zweig
Courant Institute of Mathematical Sciences
New York University
New York, NY 10012, USA
az831@nyu.edu

Joan Bruna
Courant Institute of Mathematical Sciences
Center for Data Science
New York University
New York, NY 10012, USA
bruna@cims.nyu.edu

ABSTRACT

Few neural architectures lend themselves to provable learning with gradient based methods. One popular model is the single-index model, in which labels are produced by composing an unknown linear projection with a possibly unknown scalar link function. Learning this model with SGD is relatively well-understood, whereby the so-called information exponent of the link function governs a polynomial sample complexity rate. However, extending this analysis to deeper or more complicated architectures remains challenging.

In this work, we consider single index learning in the setting of symmetric neural networks. Under analytic assumptions on the activation and maximum degree assumptions on the link function, we prove that gradient flow recovers the hidden planted direction, represented as a finitely supported vector in the feature space of power sum polynomials. We characterize a notion of information exponent adapted to our setting that controls the efficiency of learning.

1 INTRODUCTION

Quantifying the advantage of neural networks over simpler learning systems remains a primary question in deep learning theory. Specifically, understanding their ability to discover relevant low-dimensional features out of high-dimensional inputs is a particularly important topic of study. One facet of the challenge is explicitly characterizing the evolution of neural network weights through gradient-based methods, owing to the nonconvexity of the optimization landscape.

The single index setting, long studied in economics and biostatistics (Radchenko, 2015) offers the simplest setting where non-linear feature learning can be characterized explicitly. In this setting, functions of the form $x \mapsto f(\langle x, \theta^* \rangle)$ where $\theta^* \in \mathcal{S}_{d-1}$ represents a hidden direction in high-dimensional space, and f a certain non-linear link function, are learned via a student with an identical architecture $x \mapsto f(\langle x, \theta \rangle)$, under certain data distribution assumptions, such as Gaussian data. Gradient flow and gradient descent (Yehudai & Ohad, 2020; Arous et al., 2021; Dudeja & Hsu, 2018) in this setting can be analyzed by reducing the high-dimensional dynamics of θ to dimension-free dynamics of appropriate *summary statistics*, given in this case by the scalar correlation $\langle \theta, \theta^* \rangle$.

The efficiency of gradient methods in this setting, measured either in continuous time or independent samples, is controlled by two main properties. First, the correlation initialization, which typically scales as $\frac{1}{\sqrt{d}}$ for standard assumptions. Second, the information exponent s_f of f (Arous et al., 2021; Dudeja & Hsu, 2018; Bietti et al., 2022; Damian et al., 2023; 2022; Abbe et al., 2023), which measures the number of effective vanishing moments of the link function — leading to a sample complexity of the form $O(d^{s-1})$ for generic values of s.

While this basic setup has been extended along certain directions, e.g. relaxing the structure on the input data distribution Yehudai & Ohad (2020); Bruna et al. (2023), considering the multi-index counterpart Damian et al. (2022); Abbe et al. (2022; 2023); Arnaboldi et al. (2023), or learning the link function with semi-parametric methods Bietti et al. (2022); Mahankali et al. (2023), they are all fundamentally associated with fully-connected shallow neural networks. Such architecture, for all its rich mathematical structure, also comes with important shortcomings. In particular, it is unable to account for predefined symmetries in the target function that the learner wishes to exploit. This requires specialized neural architectures enforcing particular invariances, setting up novel technical challenges to carry out the program outlined above.

In this work, we consider arguably the easiest form of symmetry, given by permutation invariance. The primary architecture for this invariance is DeepSets (Zaheer et al., 2017), which is necessarily three layers by definition and

therefore not a simple extension of the two layer setting. In order to quantify the notion of 'symmetric' feature learning in this setting, we introduce a symmetric single index target, and analyze the ability of gradient descent over a DeepSets architecture to recover it. Under appropriate assumptions on the model, initialization and data distribution, we combine the previous analyses with tools from symmetric polynomial theory to characterize the dynamics of this learning problem. Our primary theorem is a proof of efficient learning under gradient flow, with explicit polynomial convergence rates controlled by an analogue of information exponent adapted to the symmetric setting. Combined with other contemporary works, this result solidifies the remarkable ability of gradient descent to perform feature learning under a variety of high-dimensional learning problems.

2 SETUP

2.1 NOTATION

For $z \in \mathbb{C}$, we will use $\overline{z}$ to denote the complex conjugate, with the notation z^* always being reserved to denote a special value of z rather than an operation. For complex matrices A we will use $A^\dagger$ to denote the conjugate transpose. The standard inner product on $\mathbb{C}^N$ is written as $\langle \cdot, \cdot \rangle$, whereas inner products on $L^2(\gamma)$ spaces for some probability measure γ will be written as $\langle \cdot, \cdot \rangle_\gamma$. Furthermore, for h a vector and $p(x)$ a vector-valued function, we will use $\langle h, p \rangle_\gamma$ as shorthand for the notation $\langle h, p(\cdot) \rangle_\gamma$.

2.2 REGRESSION SETTING AND TEACHER FUNCTION

We consider a typical regression setting, where given samples $(x, y) \in \mathcal{X} \times \mathbb{C}$ with $y = F(x)$, we seek to learn a function F_w with parameter $w \in \mathbb{C}^M$ by minimizing some expected loss $E_{x \sim \nu}\left[L(F(x), F_w(x))\right]$. Note that we consider complex-valued inputs and parameters because they greatly simplify the symmetric setting (see Proposition 2.3), hence we will also assume $\mathcal{X} \subseteq \mathbb{C}^N$. Both F and F_w will be permutation invariant functions, meaning that $F(x_{\pi(1)}, \dots x_{\pi(N)}) = F(x_1, \dots, x_N)$ for any permutation $\pi : \{1, N\} \to \{1, N\}$.

Typically the single index setting assumes that the trained architecture will exactly match the true architecture (e.g. as in Arous et al. (2021)), but below we will see why it's necessary to consider separate architectures. For that reason, we'll consider separately defining the teacher F and the student F_w.

The first ingredient are the power sum polynomials:

Definition 2.1. *For $k \in \mathbb{N}$ and $x \in \mathbb{C}^N$, the* normalized powersum polynomial *is defined as*

$$p_k(x) = \frac{1}{\sqrt{k}} \sum_{n=1}^{N} x_n^k \, .$$

Let $p(x) = [p_1(x), p_2(x), \dots]$ be an infinite dimensional vector of powersums, and consider a fixed vector $h^* \in \mathbb{C}^\infty$ of unit norm. Then our teacher function F will be of the form

$$F : \mathcal{X} \to \mathbb{C} \tag{1}$$

$$x \mapsto F(x) := f(\langle h^*, p(x) \rangle) \tag{2}$$

for some scalar link function $f : \mathbb{C} \to \mathbb{C}$. F may thus be understood as a single-index function in the feature space of powersum polynomials.

2.3 DEEPSETS STUDENT FUNCTION

Let us remind the typical structure of a DeepSets network (Zaheer et al., 2017), where for some maps $\Phi : \mathcal{X} \to \mathbb{C}^M$ and $\rho : \mathbb{C}^M \to \mathbb{C}$, the standard DeepSets architecture is of the form:

$$x \mapsto \rho\left(\Phi_1(x), \dots, \Phi_M(x)\right) \, . \tag{3}$$

The essential restriction is that Φ is a permutation invariant mapping, typically of the form $\Phi_m(x) = \sum_{n=1}^{N} \phi_m(x_n)$ for some map $\phi_m : \mathbb{C} \to \mathbb{C}$. In order to parameterize our student network as a DeepSets model, we will make the simplest possible choices, while preserving its non-linear essence. To define our student network, we consider the symmetric embedding Φ as a one-layer neural network with no bias terms:

$$\Phi_m(x) = \sum_{n=1}^{N} \sigma(a_m x_n) \, , \tag{4}$$

for i.i.d. complex weights sampled uniformly from the complex circle $a_m \sim S^1$ and some activation $\sigma : \mathbb{C} \to \mathbb{C}$. And given some link function $g : \mathbb{C} \to \mathbb{C}$, we'll consider the mapping ρ as:

$$\rho_w(\cdot) = g(\langle w, \cdot \rangle) \,, \tag{5}$$

where $w \in \mathbb{C}^M$ are our trainable weights. Putting all together, our student network thus becomes

$$\begin{aligned} F_w &: \mathcal{X} \to \mathbb{C} \\ &x \mapsto F_w(x) := g(\langle w, \Phi(x) \rangle) \,. \end{aligned} \tag{6}$$

Explicitly, this corresponds to a DeepSets network with only one trainable vector w, while the first layer weights $\{a_m\}_{m=1}^M$ and the third layer weights that parameterize the function g are frozen.

The first fact we need is that, through simple algebra, the student may be rewritten in the form of a single-index model.

Proposition 2.2. *There is matrix $A \in \mathbb{C}^{\infty \times M}$ depending only on the activation σ and the frozen weights $\{a_m\}_{m=1}^M$ such that*

$$g(\langle w, \Phi(x) \rangle) = g(\langle Aw, p(x) \rangle) \,. \tag{7}$$

2.4 Hermite-like Identity

In the vanilla single index setting, the key to giving an explicit expression for the expected loss (for Gaussian inputs) is a well-known identity of Hermite polynomials (O'Donnell, 2021; Jacobsen, 1996). If h_k denotes the Hermite polynomial of degree k, this identity takes the form

$$\langle h_k(\langle \cdot, u \rangle), h_l(\langle \cdot, v \rangle) \rangle_{\gamma_n} = \delta_{kl} k! \langle u, v \rangle^k \,, \tag{8}$$

where $u, v \in \mathbb{R}^n$ and γ_n is the standard Gaussian distribution on n dimensions.

In our setting, as it turns out, one can establish an analogous identity, by considering a different input probability measure, and a bound on the degree of the link function. We will choose our input domain $\mathcal{X} = (S^1)^N$, and the input distribution we will consider is the set of eigenvalues of a Haar-distributed unitary matrix in dimension N (Diaconis & Shahshahani, 1994), or equivalently the squared Vandermonde density over N copies of the complex unit circle (Macdonald, 1998). We'll interchangeably use the notation $\mathbb{E}_{x \sim V}[f(x)\overline{g(x)}] = \langle f, g \rangle_V$.

Proposition 2.3. *Consider $h, \tilde{h} \in \mathbb{C}^\infty$ with bounded L_2 norm. For exponents k, l with $k \leq \sqrt{N}$, if h is only supported on the first $\sqrt{N}$ elements, then:*

$$\langle \langle h, p \rangle^k, \langle \tilde{h}, p \rangle^l \rangle_V = \delta_{kl} k! \langle h, \tilde{h} \rangle^k \,. \tag{9}$$

The crucial feature of this identity is that the assumptions on support and bounded degree only apply to $\langle h, p \rangle^k$, with no restrictions on the other term. In our learning problem, we can use this property to make these assumptions on the teacher function, while requiring no bounds on the terms of the student DeepSets architecture.

In order to take advantage of the assumptions on the support of h and the degree in the above proposition, we need to make the following assumptions on our teacher link function f and our true direction h^*:

Assumption 2.4. *The link function f is analytic and only supported on the first $\sqrt{N}$ degree monomials, i.e.*

$$f(z) = \sum_{j=1}^{\sqrt{N}} \frac{\alpha_j}{\sqrt{j!}} z^j \tag{10}$$

Furthermore, the vector h^* is only supported on the first $\sqrt{N}$ elements.

Although this assumption is required to apply the orthogonality property for our loss function in the following sections, we note that in principle, including exponentially small terms of higher degree in f or higher index in h^* should have negligible effect. Moreover, one should interpret this assumption as silently disappearing in the high-dimensional regime $N \to \infty$. For simplicity, we keep this assumption to make cleaner calculations and leave the issue of these small perturbations to future work.

2.5 INFORMATION EXPONENT

Because Proposition 2.3 takes inner products of monomials, it alludes to a very simple characterization of information exponent. Namely:

Definition 2.5. *Consider an analytic function $f : \mathbb{C} \to \mathbb{C}$ that can be written in the form*

$$f(z) = \sum_{j=0}^{\infty} \frac{\alpha_j}{\sqrt{j!}} z^j \tag{11}$$

Then the information exponent *is defined as $s = \inf\{j \geq 1 : \alpha_j \neq 0\}$.*

Similar to the Gaussian case (Arous et al., 2021; Bietti et al., 2022), the information exponent s will control the efficiency of learning. Assuming $|\alpha_s|$ is some non-negligible constant, the value of s will be far more important in governing the convergence rate.

2.6 CHOOSING A LEARNABLE LOSS

There are two subtleties to choosing an appropriate loss function. Namely, the necessity of a correlational loss (with regularization), and the necessity of choosing the student and teacher link functions to be distinct.

At first glance, it is tempting to simply define a loss of the form

$$\tilde{L}(w) = \mathbb{E}_{x\sim V}|F(x) - F_w(x)|^2 = \mathbb{E}_{x\sim V}\left[|f(\langle h^*, p(x)\rangle) - f(\langle Aw, p(x)\rangle)|^2\right] . \tag{12}$$

However, the Deepsets student model is not degree limited, that is the support of Aw is not restricted to the first $\sqrt{N}$ terms of the powersum expansion. In other words, expanding this loss will require calculating the term $\|f(\langle Aw, p\rangle)\|_V^2$, which will contain high degree terms that cannot be controlled with Proposition 2.3. One could avoid this issue by choosing the activation such that Aw only contains low-index terms, but we want to consider larger classes of activations and enforce fewer restrictions.

One can instead consider a correlational loss. In this case, in order to make the objective have a bounded global minimum, it's necessary to either regularize w, or project at every step of SGD, which is the strategy taken in Damian et al. (2023). In our setting, this projection would correspond to projecting w to the ellipsoid surface $\|Aw\| = 1$. This projection would require solving an optimization problem at every timestep (Pope, 2008). To avoid this impracticality, we instead consider regularization.

Then with complete knowledge of the link function f, specifically its monomial coefficients, we can now define the correlational loss

$$\hat{L}(w) = \mathbb{E}_{x\sim V}\left[-\operatorname{Re}\left\{f(\langle h^*, p(x)\rangle \overline{f(\langle Aw, p(x)\rangle)}\right\}\right] + \sum_{i=j}^{\sqrt{N}} \frac{|\alpha_j|^2}{2}\|Aw\|^{2j} . \tag{13}$$

This loss enjoys benign optimization properties, as shown by the following proposition:

Proposition 2.6. *If there exist coprimes k, l with $\alpha_k, \alpha_l \neq 0$, and h^* is in the range of A, then $\hat{L}$ exclusively has global minima at all w such that $Aw = h^*$.*

However, unlike the real case, complex weights causes issues for learning this objective. Namely, this objective can be written as a non-convex polynomial in $\cos\theta$ where θ is the angle of $\langle Aw, h^*\rangle$ in polar coordinates.

Therefore, we consider a different choice of student link function that will enable a simpler analysis of the dynamics. For the choice of $g(z) = \frac{\alpha_s}{|\alpha_s|\sqrt{s!}} z^s$, we instead consider the loss:

$$L(w) = \mathbb{E}_{x\sim V}\left[-\operatorname{Re}\left\{f(\langle h^*, p(x)\rangle \overline{g(\langle Aw, p(x)\rangle)}\right\}\right] + \frac{|\alpha_s|}{2}\|Aw\|^{2s} \tag{14}$$

$$= -|\alpha_s| \operatorname{Re}\{\langle Aw, h^*\rangle^s\} + \frac{|\alpha_s|}{2}\|Aw\|^{2s} . \tag{15}$$

We note that Dudeja & Hsu (2018) used a similar trick of a correlational loss containing a single orthogonal polynomial in order to simplify the learning landscape. The global minima of this loss, and in fact the dynamics of gradient flow on it, will be explored in the sequel.

3 Related Work

3.1 Single Index Learning

The conditions under which single-index model learning is possible have been well-explored in previous literature. The main assumptions that enable provably learning under gradient flow / gradient descent are monotonicity of the link function (Kakade et al., 2011; Kalai & Sastry, 2009; Shalev-Shwartz et al., 2010; Yehudai & Ohad, 2020) or Gaussian input distribution (Arous et al., 2021). The former assumption essentially corresponds to the setting where the information exponent $s = 1$, as it will have positive correlation with a linear term. Under the latter assumption, the optimal sample complexity was achieved in Damian et al. (2023), with study of learning when the link function is not known in Bietti et al. (2022).

When both assumptions are broken, the conditions on the input distribution of rotation invariance or approximate Gaussianity are nevertheless sufficient for learning guarantees (Bruna et al., 2023). But more unusual distributions, especially in the complex domain that is most convenient for symmetric networks, are not well studied.

3.2 Symmetric Neural Networks

The primary model for symmetric neural networks was introduced in Zaheer et al. (2017) as the DeepSets model. There are many similar models that enforce permutation invariance (Qi et al., 2017; Santoro et al., 2017; Lee et al., 2019), though we focus on DeepSets because of its relationship with the power sum polynomials and orthogonality (Zweig & Bruna, 2022). We are not aware of any other works that demonstrate provable learning of symmetric functions under gradient-based methods.

4 Provably Efficient Recovery with Gradient Flow

4.1 Defining the Dynamics

The gradient methods considered in Arous et al. (2021); Ben Arous et al. (2022) are analyzed by reducing to a dimension-free dynamical system of the so-called summary statistics. For instance, in the vanilla single-index model, the summary statistics reduce to the scalar correlation between the learned weight and the true weight. In our case, we have three variables, owing to the fact that the correlation is complex and represented by two scalars, and a third variable controlling the norm of the weight since we aren't using projection.

Note that although our weight vector w is complex, we still apply regular gradient flow to the pair of weight vectors w_R, w_C where $w = w_R + iw_C$. Furthermore, we use the notation $\nabla := \nabla_w = \nabla_{w_R} + i\nabla_{w_C}$. With that in mind, we can summarize the dynamics of our gradient flow in the following Theorem.

Theorem 4.1. *Given a parameter w, consider the summary statistics $m = \langle Aw, h^* \rangle \in \mathbb{C}$ and $v = \|P_{h^*}^{\perp} Aw\|^2$ where $P_{h^*}^{\perp}$ is projection onto the orthogonal complement of h^*. Let the polar decomposition of m be $re^{i\theta}$.*

Then given the preconditioned gradient flow given by

$$\dot{w} = -\frac{1}{s|\alpha_s|}(A^\dagger A)^{-1}\nabla L(w) \,, \tag{16}$$

the summary statistics obey the following system of ordinary differential equations:

$$\dot{r} = (1-\delta)r^{s-1}\cos s\theta - (v + r^2)^{s-1} r \,, \tag{17}$$

$$\frac{d}{dt}\cos s\theta = (1-\delta)sr^{s-2}(1 - \cos^2 s\theta) \,, \tag{18}$$

$$\dot{v} = 2\delta r^s \cos s\theta - 2(v + r^2)^{s-1} v \,, \tag{19}$$

where $\delta := 1 - \|P_A h^*\|^2$ and P_A is the projection onto the range of A.

The proof is in Appendix D. The main technical details come from using Wirtinger calculus to determine how the real and imaginary parts of w evolve under the flow. Additionally, the correct preconditioner (intuitive from the linear transform of w) is crucial for reducing the dynamics to only three summary statistics, and converting to dynamics on $\cos s\theta$ rather than θ itself simplifies the description of the learning in the next section dramatically.

4.2 PROVABLE LEARNING

These dynamics naturally motivate the question of learning efficiency, measured in convergence rates in time in the case of gradient flow. Our main result is that, under some assumptions on the initialization of the frozen weights $\{a_m\}_{m=1}^M$ and the initialized weight vector w_0, the efficiency is controlled by the initial correlation with the true direction and the information exponent, just as in the Gaussian case.

Theorem 4.2. *Consider a fixed $\epsilon > 0$. Suppose the initialization of w_0 and $(a_m)_{m=1}^M$ are such that:*

(i) Small correlation and anti-concentration at initialization: $0 < r_0 \leq 1$,

(ii) Initial phase condition: $\cos s\theta_0 \geq 1/2$,

(iii) Initial magnitude condition for Aw: $\|Aw_0\| = 1$, or equivalently $v_0 = 1 - r_0^2$,

(iv) Small Approximation of optimal error: $\delta \leq \min(\epsilon/2, O(s^{-s} r_0^4))$.

Then if we run the gradient flow given in Theorem 4.1 we have ϵ accuracy in the sense that:

$$r_T \geq 1 - \epsilon \,, \;\; \cos s\theta_T \geq 1 - \epsilon \,, \;\; v_T \leq \epsilon \tag{20}$$

after time T, where depending on the information exponent s:

$$T \leq \begin{cases} O\left(\log \frac{1}{\epsilon}\right) & s = 1 \,, \\ O\left(2^{s^2} r_0^{-4s} + \log \frac{1}{\epsilon}\right) & s > 1 \,. \end{cases} \tag{21}$$

Remark 4.3. *We note that we only recover $\cos s\theta \approx 1$, rather than a guarantee that $\theta \approx 0$, and so the hidden direction is only determined up to scaling by a sth root of unity. This limitation is may appear to be an issue with the choice of the student link function g, but it is unavoidable: if the teacher link function $f(z) = \frac{1}{\sqrt{s!}} z^s$, one can calculate that for any choice of g, $L(w)$ is invariant to scaling w by an sth root of unity.*

4.3 INITIALIZATION GUARANTEES

In order to apply the gradient flow bound proved in Theorem 4.2, it only remains to understand when the assumptions on initialization are met. Unlike the single-index setting with Gaussian inputs, the initial correlation is not guaranteed to be on the scale of $\frac{1}{\sqrt{N}}$, but will depend on the activation function and the random weights in the first layer. Let us introduce the assumptions we'll need:

Assumption 4.4. *We assume an analytic activation $\sigma(z) = \sum_{k=0}^{\infty} c_k z^k$, with the notation $\sigma_+ := \max_{1 \leq k \leq N} |c_k| \sqrt{k}$ and $\sigma_- := \min_{1 \leq k \leq \sqrt{N}} |c_k| \sqrt{k}$. We further assume:*

(i) $c_k = 0$ iff $k = 0$,

(ii) σ analytic on the unit disk,

(iii) $1/\sigma_- = O(\mathrm{poly}(N))$,

(iv) $\sum_{k=N+1}^{\infty} k|c_k|^2 \leq e^{-\Omega(\sqrt{N})}$.

The first two conditions are simply required for the application of Proposition 2.3, as the powersum vector p is built out of polynomials induced by the activation and does not include a constant term. The second two conditions concern the decay of the coefficients of σ, in the sense that the decay must start slow but eventually become very rapid. These conditions are necessary mainly for ensuring the Small Approximation of optimal error condition:

Lemma 4.5. *Let σ satisfy Assumption 4.4, and assume $M = O(N^3)$. Consider any unit norm $h^* \in \mathbb{C}^\infty$ that is only supported on the first $\sqrt{N}$ elements. If we sample the weights i.i.d. uniformly on the complex unit circle, $a_m \sim S^1$, with probability $1 - 2\exp(-\Omega(N))$:*

$$1 - \|P_A h^*\|^2 \leq e^{-\Omega(\sqrt{N})} \,.$$

Lastly, we can choose an initialization scheme for w which handily ensures the remaining assumptions we need to apply Theorem 4.2. The crucial features of σ are similar to the previous result. Namely, we want the initial correlation r_0 to be non-negligible because this directly controls the runtime of gradient flow. Slow initial decay with fast late decay of the σ coefficients directly implies that Aw_0 has a lot of mass in the first $\sqrt{N}$ indices and very little mass past the first N indices. These requirements rule out, say, $\exp$ as an analytic activation because the coefficients decay too rapidly.

Lemma 4.6. *Suppose w is sampled from a standard complex Gaussian on M variables. It follows that if we set $w_0 = \frac{w}{\|Aw\|}$, and use the summary statistics from Theorem 4.1, then with probability $1/3 - 2\exp(-\Omega(N))$ and any h^* as in Lemma 4.5*

(i) $1 \geq r_0 \geq c\frac{\sigma_-}{\sigma_+\sqrt{M}}$ *for some universal constant* $c > 0$,

(ii) $\cos s\theta_0 \geq 1/2$,

(iii) $v_0 = 1 - r_0^2$.

Finally, we consider a straightforward choice of σ that meets Assumption 4.4 so that we can arrive at an explicit complexity bound on learning:

Corollary 4.7 (Non-asymptotic Rates for Gradient Flow)**.** *Consider $\xi = 1 - \frac{1}{\sqrt{N}}$ and the specific choice of activation*

$$\sigma(z) = \arctan \xi z + \xi z \arctan \xi z \; .$$

Suppose we initialize w from a standard complex Gaussian in dimension M with $M = O(N^3)$, and $\{a_m\}_{m=1}^M \sim S^1$ iid. Furthermore, treat s and ϵ as constants relative to N. Then with probability $1/3 - 2\exp(-\Omega(N))$, we will recover ϵ accuracy in time

$$T \leq \begin{cases} O\left(\log \frac{1}{\epsilon}\right) & s = 1 \\ O\left(2^{s^2} N^{7s} + \log \frac{1}{\epsilon}\right) & s > 1 \; . \end{cases} \tag{22}$$

Proof. By Proposition H.5, the activation σ given in the corollary statement satisfies Assumption 4.4, so we can apply Lemma 4.6 and Lemma 4.5 to satisfy the requirements of Theorem 4.2. In particular, the fourth condition is given by assuming $e^{-\Omega(\sqrt{N})} \leq \min(\epsilon/2, O(s^{-s} r_0^4))$ which is true when s is constant, and ϵ and r_0 are at most polynomial compared to N.

Note that $\sigma_+ = O(1)$ and $\sigma_- = \Omega\left(\frac{1}{N^{1/4}}\right)$, so it follows that $r_0 = \Omega\left(\frac{1}{N^{7/4}}\right)$ with probability $1/3 - 2\exp(-\Omega(N))$. Conditioning on this bound gives the desired bound on the time for ϵ accuracy. □

Hence, we have a rate that, for $s = O(1)$, is not cursed by dimensionality to recover the true hidden direction h^*. As mentioned above, there are two caveats to this recovery: w is only recovered up to an sth root of unity, and to directly make predictions of the teacher model would require using the teacher link function rather than using the student model directly.

Since this result concerns gradient flow over the population loss, a natural question is what barriers exist that stymie the SGD analysis of recent single index papers (Arous et al., 2021; Damian et al., 2023; Bruna et al., 2023). These works treat the convergence of SGD by a standard drift and martingale argument, where the drift follows the population gradient flow, and the martingales are shown to be controlled via standard concentration inequalities and careful arguments around stopping times. Applying these tactics to a discretized version of the dynamics given in Theorem 4.1 mainly runs into an issue during the first phase of training. Unlike in Arous et al. (2021) where the drift dynamics have the correlation monotonically increasing towards 1, at the start of our dynamics the correlation magnitude r and the "orthogonal" part of the learned parameter v are both decreasing (with high probability over the initialization). Showing that this behavior doesn't draw the model towards the saddle point where $r = 0$ requires showing that v decreases meaningfully faster than r, i.e. showing that $\frac{d}{dt} \log \frac{r^2}{v}$ is positive. It's not clear what quality of bounds the martingale concentration inequalities would provide for this quantity, and we leave for future work if the six stage proof of the dynamics behavior could be successfully discretized.

5 Experiments

To study an experimental setup for our setting, we consider the student-teacher setup outlined above with gradient descent. We consider $N = 25$, $M = 100$, and approximate the matrix A by capping the infinite number of rows at 150, which was sufficient for $1 - \|P_A h^*\|^2 \leq 0.001$ in numerical experiments. For the link function f, we choose its only non-zero monomial coefficients to be $\alpha_3 = \alpha_4 = \alpha_5 = \frac{1}{\sqrt{3}}$. And correspondingly, g simply has $\alpha_3 = 1$ and all other coefficients at zero.

We choose for convenience an activation function such that $A_{km} = \left(\frac{N-1}{N}\right)^k a_m^k$. We make this choice because, while obeying all the assumptions required in Assumption 4.4, this choice implies that the action of A on the elementary

basis vectors e_j for $1 \leq j \leq \sqrt{N}$ is approximately distributed the same. This choice means that $\|P_A h^*\|$ is less dependent on the choice of h^*, and therefore reduces the variance in our experiments when we choose h^* uniformly among unit norm vectors with support on the first $\sqrt{N}$ elements, i.e. uniformly from the complex sphere in degree $\sqrt{N}$.

Under this setup, we train full gradient descent on 50000 samples from the Vandermonde V distribution under 20000 iterations. The only parameter to be tuned is the learning rate, and we observe over the small grid of $[0.001, 0.0025, 0.005]$ that a learning rate of 0.0025 performs best for the both models in terms of probability of r reaching approximately 1, i.e. strong recovery.

As described in Theorem 4.1, we use preconditioned gradient descent using $(A^\dagger A)^{-1}$ as the preconditioner, which can be calculated once at the beginning of the algorithm and is an easy alteration to vanilla gradient descent to implement. We use the pseudoinverse for improved stability in calculating this matrix, although we note that this preconditioner doesn't introduce stability issues into the updates of our summary statistics, even in the case of gradient descent. Indeed, even if one considers the loss $L(w)$ under an empirical expectation rather than full expectation, the gradient $\nabla L(w)$ can still be seen to be written in the form $A\dagger v$ for some vector v. If one preconditions this gradient by $(A^\dagger A)^{-1}$, and observes that the summary statistics m and v both depend on Aw rather than w directly, it follows that the gradient update on these statistics is always of the form $A(A^\dagger A)^{-1}A^\dagger = P_A$, so even in the empirical case this preconditioner doesn't introduce exploding gradients.

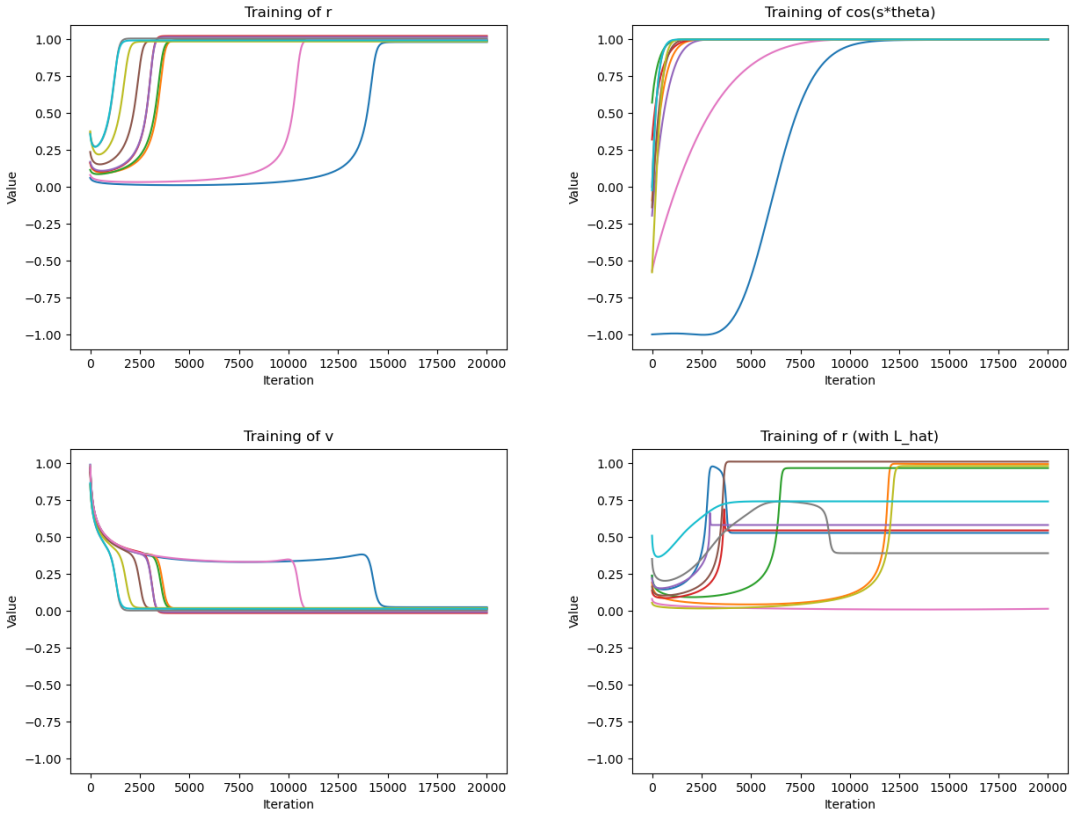

Figure 1: The learning trajectory, over ten independent runs, of the three summary statistics in the case of our chosen loss function L, and the trajectory of the r statistic for the more complicated loss function $\hat{L}$

6 Discussion

6.1 Experimental Results

The outcomes of our experiments are given in Figure 1. We observe very high rates of strong recovery using the loss L. For the loss $\hat{L}$, we note that r often becomes stuck, indicating the model has reached a local minima.

We note that our analysis is somewhat pessimistic, as the experimental gradient descent on $L(w)$ will often achieve near perfect accuracy even if $\cos s\theta_0 < 0$. This is mainly an issue of proof technique: although $\cos s\theta$ is always increasing under the dynamics, r is necessarily decreasing for as long as $\cos s\theta$ is negative. It is quite subtle to control whether $\cos s\theta$ will become positive before r becomes extremely small, and the initialization of r is the main feature that controls the runtime of the model. However the empirical results suggest that a chance of success $> 1/2$ is possible under a more delicate analysis.

However, the analysis given in the proof of Theorem 4.2 does accurately capture the brief dip in the value of r in the initial part of training, when the regularization contributes more to the gradient than the correlation until $\cos s\theta$ becomes positive.

Because we can only run experiments on gradient descent rather than gradient flow, we observe the phenomenon of search vs descent studied in Arous et al. (2021), where the increase in the correlation term r is very slow and then abruptly increases. For the model trained with $\hat{L}$, we observe that there is much greater likelihood of failure in the recovery, as r appears to become stuck below the optimal value of 1.

6.2 Extensions

The success of this method of analysis depends heavily on the Hermite-like identity in Proposition 2.3. In general, many of the existing results analyzing single index models need to assume either Gaussian inputs, or uniformly distributed inputs on the Boolean hypercube (see for example Abbe et al. (2023)). In some sense, this works cements the inclusion of the Vandermonde distribution in this set of measures that enable clean analysis. The proof techniques for these three measures are quite disparate, so it remains open to determine if there is a wider class of "nice" distributions where gradient dynamics can be successfully analyzed.

Additionally, the success of the multi-layer training in Bietti et al. (2022); Mahankali et al. (2023) suggests that simultaneously training the frozen first layer weights may not prohibit the convergence analysis. The matrix A depends on the first layer weights through a Vandermonde matrix (see X in the proof of Lemma 4.5), and the simple characterization of the derivative of a Vandermonde matrix alludes to further possibilities for clean analysis.

6.3 Limitations

A first limitation is the focus of this work on complex inputs, analytic activations, and fixed input distribution (namely the squared Vandermonde density). Although complex analytic functions are less commonly studied in the literature, they do still appear in settings like quantum chemistry (Beau & del Campo, 2021; Langmann, 2005). Regarding the focus on the Vandermonde distribution, we note this is similar to the vanilla single-index setting in the restriction to Gaussian inputs, under which the theory is particularly powerful, simplest and understanding of non-Gaussian data is still nascent.

A second limitation is that this work focuses on input distributions over sets of scalars, whereas typically symmetric neural networks are applied to sets of high-dimensional vectors. Proposition 2.3 does not work out of the box for these settings without a high-dimensional analogue of the inner product $\langle \cdot, \cdot \rangle_V$ with similar orthogonality properties. It is possible to define such an inner products on the so-called multisymmetric powersums with similar orthogonality (Zweig & Bruna, 2022), and we leave to future work the question of whether such inner products could grant similar guarantees about the learning dynamics in this more realistic setting.

7 Conclusion

In this work we've shown a first positive result that quantifies the ability of gradient descent to perform symmetric feature learning, by adapting and extending the tools of two-layer single index models. In essence, this is made possible by a 'miracle', namely the fact that certain powersum expansions under the Vandermonde measure enjoy the same semigroup structure as Hermite polynomials under the Gaussian measure (Proposition 2.3) — leading to a dimension-free summary statistic representation of the loss. Although the resulting dynamics are more intricate than in the Euclidean setting, we are nonetheless able to establish quantitative convergence rates to 'escape the mediocrity' of initialization, recovering the same main ingredients as in previous works Arous et al. (2021); Abbe et al. (2022), driven by the information exponent. To our knowledge, this is the first work to show how learning with gradient based methods necessarily succeeds in this fully non-linear (i.e. not in the NTK regime) setting. Nevertheless, there are many lingering questions.

As discussed, one limitation of the analysis is the reliance on gradient flow rather than gradient descent. We hope that in future work we'll be able to effectively discretize the dynamics, made more challenging by the fact that one must track three parameters rather than simply the correlation. Still, we observe theoretically and empirically that the symmetric single index setting demands a number of unusual choices, such as a correlation loss and distinct student and teacher link function, in order to enable efficient learning. And in a broader scheme, if one remembers the perspective of DeepSets as a very limited form of a three-layer architecture, the issue of provable learning for deeper, more realistic architectures stands as a very important and unexplored research direction — and where Transformers with planted low-dimensional structures appear as the next natural question.

REFERENCES

Emmanuel Abbe, Enric Boix-Adsera, and Theodor Misiakiewicz. The merged-staircase property: a necessary and nearly sufficient condition for sgd learning of sparse functions on two-layer neural networks. *arXiv preprint arXiv:2202.08658*, 2022.

Emmanuel Abbe, Enric Boix-Adsera, and Theodor Misiakiewicz. Sgd learning on neural networks: leap complexity and saddle-to-saddle dynamics. *arXiv preprint arXiv:2302.11055*, 2023.

Luca Arnaboldi, Ludovic Stephan, Florent Krzakala, and Bruno Loureiro. From high-dimensional & mean-field dynamics to dimensionless odes: A unifying approach to sgd in two-layers networks. *arXiv preprint arXiv:2302.05882*, 2023.

Gerard Ben Arous, Reza Gheissari, and Aukosh Jagannath. Online stochastic gradient descent on non-convex losses from high-dimensional inference. *The Journal of Machine Learning Research*, 22(1):4788–4838, 2021.

Mathieu Beau and Adolfo del Campo. Parent hamiltonians of jastrow wavefunctions. *SciPost Physics Core*, 4(4):030, 2021.

Gerard Ben Arous, Reza Gheissari, and Aukosh Jagannath. High-dimensional limit theorems for sgd: Effective dynamics and critical scaling. *arXiv preprint arXiv:2206.04030*, 2022.

Alberto Bietti, Joan Bruna, Clayton Sanford, and Min Jae Song. Learning single-index models with shallow neural networks. In *Advances in Neural Information Processing Systems*, 2022.

Imre Bihari. A generalization of a lemma of bellman and its application to uniqueness problems of differential equations. *Acta Mathematica Hungarica*, 7(1):81–94, 1956.

Joan Bruna, Loucas Pillaud-Vivien, and Aaron Zweig. On single index models beyond gaussian data. *arXiv preprint arXiv:2307.15804*, 2023.

Alex Damian, Eshaan Nichani, Rong Ge, and Jason D Lee. Smoothing the landscape boosts the signal for sgd: Optimal sample complexity for learning single index models. *arXiv preprint arXiv:2305.10633*, 2023.

Alexandru Damian, Jason Lee, and Mahdi Soltanolkotabi. Neural networks can learn representations with gradient descent. In *Conference on Learning Theory*, 2022.

Persi Diaconis and Mehrdad Shahshahani. On the eigenvalues of random matrices. *Journal of Applied Probability*, 31 (A):49–62, 1994.

Rishabh Dudeja and Daniel Hsu. Learning single-index models in gaussian space. In *Conference On Learning Theory*, pp. 1887–1930. PMLR, 2018.

Robert FH Fischer. *Precoding and signal shaping for digital transmission*. John Wiley & Sons, 2005.

Martin Jacobsen. Laplace and the origin of the ornstein-uhlenbeck process. *Bernoulli*, 2(3):271–286, 1996.

Sham M Kakade, Varun Kanade, Ohad Shamir, and Adam Kalai. Efficient learning of generalized linear and single index models with isotonic regression. *Advances in Neural Information Processing Systems*, 24, 2011.

Adam Tauman Kalai and Ravi Sastry. The isotron algorithm: High-dimensional isotonic regression. In *COLT*, 2009.

Edwin Langmann. A method to derive explicit formulas for an elliptic generalization of the jack polynomials. *arXiv preprint math-ph/0511015*, 2005.

Juho Lee, Yoonho Lee, Jungtaek Kim, Adam Kosiorek, Seungjin Choi, and Yee Whye Teh. Set transformer: A framework for attention-based permutation-invariant neural networks. In *International Conference on Machine Learning*, pp. 3744–3753. PMLR, 2019.

Ian Grant Macdonald. *Symmetric functions and Hall polynomials*. Oxford university press, 1998.

Arvind Mahankali, Jeff Z Haochen, Kefan Dong, Margalit Glasgow, and Tengyu Ma. Beyond ntk with vanilla gradient descent: A mean-field analysis of neural networks with polynomial width, samples, and time. *arXiv preprint arXiv:2306.16361*, 2023.

Ryan O'Donnell. Analysis of boolean functions. *arXiv preprint arXiv:2105.10386*, 2021.

Stephen B Pope. Algorithms for ellipsoids. *Cornell University Report No. FDA*, pp. 08–01, 2008.

Charles R Qi, Hao Su, Kaichun Mo, and Leonidas J Guibas. Pointnet: Deep learning on point sets for 3d classification and segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 652–660, 2017.

Peter Radchenko. High dimensional single index models. *Journal of Multivariate Analysis*, 139:266–282, 2015.

Adam Santoro, David Raposo, David G Barrett, Mateusz Malinowski, Razvan Pascanu, Peter Battaglia, and Timothy Lillicrap. A simple neural network module for relational reasoning. *Advances in neural information processing systems*, 30, 2017.

Shai Shalev-Shwartz, Ohad Shamir, and Karthik Sridharan. Learning kernel-based halfspaces with the zero-one loss. *arXiv preprint arXiv:1005.3681*, 2010.

Roman Vershynin. *High-dimensional probability: An introduction with applications in data science*, volume 47. Cambridge university press, 2018.

Gilad Yehudai and Shamir Ohad. Learning a single neuron with gradient methods. In *Conference on Learning Theory*, pp. 3756–3786. PMLR, 2020.

Manzil Zaheer, Satwik Kottur, Siamak Ravanbakhsh, Barnabas Poczos, Russ R Salakhutdinov, and Alexander J Smola. Deep sets. *Advances in neural information processing systems*, 30, 2017.

Aaron Zweig and Joan Bruna. Exponential separations in symmetric neural networks. *Advances in Neural Information Processing Systems*, 35:33134–33145, 2022.

A PROOF OF PROPOSITION 2.2

The proposition is true for any activation with a Laurent series, but we will only prove it for activations satisfying Assumption 4.4 since that's the only setting we'll require it.

Consider an analytic activation σ with no constant term, given as

$$\sigma(z) = \sum_{k=1}^{\infty} c_k z^k \tag{23}$$

And remind the deepsets features map

$$\Phi_m(x) = \sum_{n=1}^{N} \sigma(a_m x_n) \tag{24}$$

$$\tag{25}$$

where we have neurons without bias terms.

Then for a weight $w \in \mathbb{C}^M$, one can quickly see

$$\begin{aligned}
\langle w, \Phi(x) \rangle &= \sum_{m=1}^{M} w_m \sum_{n=1}^{N} \sigma(a_m x_n) && (26)\\
&= \sum_{m=1}^{M} w_m \sum_{n=1}^{N} \sum_{k=1}^{\infty} c_k (a_m x_n)^k && (27)\\
&= \sum_{k=1}^{\infty} \sum_{m=1}^{M} w_m c_k a_m^k \sqrt{k} p_k(x) && (28)\\
&= \langle Aw, p(x) \rangle && (29)
\end{aligned}$$

where $A_{km} = c_k \sqrt{k} a_m^k$

B PROOF OF PROPOSITION 2.3

We require some definitions to use the machinery of symmetric polynomials.

Definition B.1. *An* integer partition λ *is non-increasing, finite sequence of positive integers* $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_k$. *The weight of the partition is given by* $|\lambda| = \sum_{i=1}^{k} \lambda_i$. *The length of a partition* $l(\lambda)$ *is the number of terms in the sequence.*

Then we characterize a product of powersums by:

$$p_\lambda(x) = \prod_i p_{\lambda_i}(x) \tag{30}$$

Finally, define the combinatorial constant $t_\lambda = \prod_{i=1}^{|\lambda|} (m_i)!$ where m_i denotes the number of parts of λ equal to i.

Theorem B.2 ((Macdonald, 1998, Chapter VI (9.10)))**.** *For partitions* λ, μ *with* $|\lambda| \leq N$*:*

$$\langle p_\lambda, p_\mu \rangle_V = t_\lambda \mathbb{1}_{\lambda=\mu} \tag{31}$$

With that in mind, let's consider the inner product of two simple single-index functions.

Let $p = [p_1, p_2, \dots]$ be an infinite vector of powersums, and choose exponents i, j with $i \leq \sqrt{N}$. Then for any $h, \tilde{h} \in \mathbb{C}^\infty$ such that h is only supported on the first $\sqrt{N}$ entries:

$$\langle\langle h,p\rangle^i,\langle\tilde{h},p\rangle^j\rangle_V = \left\langle \sum_{|\alpha|=i}\binom{i}{\alpha}h^\alpha\overline{p^\alpha}, \sum_{|\alpha|=j}\binom{j}{\alpha}\tilde{h}^\alpha\overline{p^\alpha},\right\rangle_V \tag{32}$$

$$= \delta_{ij}\sum_{|\alpha|=i}\binom{i}{\alpha}^2\langle p^\alpha,p^\alpha\rangle_V h^\alpha\overline{\tilde{h}^\alpha} \tag{33}$$

$$= \delta_{ij}\sum_{|\alpha|=i}\binom{i}{\alpha}^2\left(\prod_{k=1}^{\sqrt{N}}\alpha_k!\right)h^\alpha\overline{\tilde{h}^\alpha} \tag{34}$$

$$= \delta_{ij}i!\sum_{|\alpha|=i}\binom{i}{\alpha}h^\alpha\overline{\tilde{h}^\alpha} \tag{35}$$

$$= \delta_{ij}i!\langle h,\tilde{h}\rangle^i \tag{36}$$

C Proof of Propsition 2.6

Applying Proposition 2.3 and using assumptions on the degree bound on f and the support of h^*, we can write:

$$\langle f(\langle h^*,p\rangle), f(\langle Aw,p\rangle)\rangle_V = \left\langle \sum_{j=1}^{\sqrt{N}}\frac{\alpha_j}{\sqrt{j!}}\langle h^*,p\rangle^j, \sum_{j=1}^{\sqrt{N}}\frac{\alpha_j}{\sqrt{j!}}\langle Aw,p\rangle^j\right\rangle_V \tag{37}$$

$$= \sum_{j=1}^{\sqrt{N}}|\alpha_j|^2\langle h^*,Aw\rangle^j \tag{38}$$

Hence we have

$$\hat{L}(w) = E_{x\sim V}\left[-\operatorname{Re}\left\{f(\langle h^*,p(x)\rangle\overline{f(\langle Aw,p(x)\rangle)}\right\}\right] + \sum_{j=1}^{\sqrt{N}}\frac{|\alpha_j|^2}{2}\|Aw\|^{2j} \tag{39}$$

$$= -\frac{1}{2}\langle f(\langle h^*,p\rangle), f(\langle Aw,p\rangle)\rangle_V - \frac{1}{2}\overline{\langle f(\langle h^*,p\rangle), f(\langle Aw,p\rangle)\rangle_V} + \sum_{j=1}^{\sqrt{N}}\frac{|\alpha_j|^2}{2}\|Aw\|^{2j} \tag{40}$$

$$= -\sum_{j=1}^{\sqrt{N}}|\alpha_j|^2\operatorname{Re}\{\langle h^*,Aw\rangle^j\} + \frac{|\alpha_j|^2}{2}\|Aw\|^{2j} \tag{41}$$

Now, we use the same notation as in Theorem 4.1 and introduce variables $m = \langle Aw,h^*\rangle = re^{i\theta}$ and $v = \|Aw\|^2 - r^2$, such that we can write:

$$\hat{L}(w) = \sum_{j=1}^{\sqrt{N}}|\alpha_j|^2\left(-r^j\cos j\theta + \frac{1}{2}(v+r^2)^j\right) \tag{42}$$

Because $r \geq 0$ and $v \geq 0$, this loss can be minimized by setting $v = 0$ and any θ where $\cos j\theta = 1$ for all j with $\alpha_j \neq 0$. Since we assume there are distinct indices i,j that are coprime with non-zero support, we require $i\theta$ and $j\theta$ to both be multiples of 2π, which is only possible if $\theta \equiv 0 \mod 2\pi$. Therefore:

$$\hat{L}(w) = \sum_{j=1}^{\sqrt{N}} |\alpha_j|^2 \left(-r^j + \frac{1}{2} r^{2j} \right) \tag{43}$$

$$= C + \sum_{j=1}^{\sqrt{N}} \frac{|\alpha_j|^2}{2} \left(r^j - 1 \right)^2 \tag{44}$$

for some constant C, and this is minimized at $r = 1$. Hence, if $r = 1, \theta \equiv 0 \mod 2\pi, v = 0$, it follows that $Aw = h^*$.

D Proof of Theorem 4.1

Given the matrix A and weight w, an identical calculation to the one in Proposition 2.6 shows that

$$L(w) = E_{x \sim V} \left[-\operatorname{Re}\left\{ f(\langle h^*, p(x) \rangle \overline{g(\langle Aw, p(x) \rangle)} \right\} \right] + \frac{|\alpha_s|}{2} \|Aw\|^{2s} \tag{45}$$

$$= -|\alpha_s| \operatorname{Re}\{\langle Aw, h^* \rangle^s\} + \frac{|\alpha_s|}{2} \|Aw\|^{2s} \tag{46}$$

To calculate the gradient with respect to the real and imaginary parts of w, we use tools from Wirtinger calculus (Fischer, 2005). Using the notation that $\nabla_{\overline{w}} = \frac{1}{2}(\nabla_{w_R} + i \nabla_{w_C})$ and the appropriate generalization of the chain rule, we have:

$$2\nabla_{\overline{w}} \operatorname{Re}\{\langle Aw, h^* \rangle^s\} = \nabla_{\overline{w}} \left(\langle Aw, h^* \rangle^s + \overline{\langle Aw, h^* \rangle^s} \right) \tag{47}$$

$$= \nabla_{\overline{w}} \overline{\langle Aw, h^* \rangle}^s \tag{48}$$

$$= s \overline{\langle Aw, h^* \rangle}^{s-1} A^\dagger h^* \tag{49}$$

Likewise,

$$2\nabla_{\overline{w}} \|Aw\|^{2s} = 2s \|Aw\|^{2(s-1)} \nabla_{\overline{w}} \|Aw\|^2 \tag{50}$$

$$= 2s \|Aw\|^{2(s-1)} \nabla_{\overline{w}} \left(w^\dagger A^\dagger A w \right) \tag{51}$$

$$= 2s \|Aw\|^{2(s-1)} A^\dagger A w \tag{52}$$

Thus, we have:

$$\nabla L = \nabla_{w_R} L + i \nabla_{w_C} L \tag{53}$$

$$= 2\nabla_{\overline{w}} L \tag{54}$$

$$= -s|\alpha_s| \overline{\langle Aw, h^* \rangle}^{s-1} A^\dagger h^* + s|\alpha_s| \|Aw\|^{2(s-1)} A^\dagger A w \tag{55}$$

We introduce the parameters

$$m = \langle Aw, h^* \rangle = \langle w, A^\dagger h^* \rangle \tag{56}$$

$$v = \|P_{h^*}^\perp Aw\|^2 = \|Aw\|^2 - |m|^2 \tag{57}$$

And we consider preconditioned gradient flow of the form (where for complex variables we use similar notation that $\dot{w} = \dot{w_R} + \dot{w_C} i$):

$$\dot{w} = -\frac{1}{s|\alpha_s|}(A^\dagger A)^{-1}\nabla L \tag{58}$$
$$= \overline{m}^{s-1}(A^\dagger A)^{-1}A^\dagger h^* - \|Aw\|^{2(s-1)}w \tag{59}$$

It follows that

$$\dot{m} = \langle \dot{w}, A^\dagger h^* \rangle \tag{60}$$
$$= \|P_A h^*\|^2 \overline{m}^{s-1} - (v + |m|^2)^{s-1} m \tag{61}$$

where $P_A = A(A^\dagger A)^{-1}A^\dagger$ is the orthogonal projection onto the range of A.

Let $m = a + bi = re^{i\theta}$, so we have $\dot{m} = \dot{a} + \dot{b}i$. Thus

$$\dot{a} = \|P_A h^*\|^2 r^{s-1}\cos(s-1)\theta - (v + r^2)^{s-1} r\cos\theta \tag{62}$$
$$\dot{b} = -\|P_A h^*\|^2 r^{s-1}\sin(s-1)\theta - (v + r^2)^{s-1} r\sin\theta \tag{63}$$

Now we do a change of variables, because $a = r\cos\theta$ and $b = r\sin\theta$, so

$$\dot{a} = \dot{r}\cos\theta - r\dot{\theta}\sin\theta \tag{64}$$
$$\dot{b} = \dot{r}\sin\theta + r\dot{\theta}\cos\theta \tag{65}$$
$$\tag{66}$$

Rearranging, we can get the flow on r and θ:

$$\dot{r} = \dot{a}\cos\theta + \dot{b}\sin\theta \tag{67}$$
$$= \|P_A h^*\|^2 r^{s-1}\cos s\theta - (v + r^2)^{s-1} r \tag{68}$$
$$r\dot{\theta} = -\dot{a}\sin\theta + \dot{b}\cos\theta \tag{69}$$
$$= -\|P_A h^*\|^2 r^{s-1}\sin s\theta \tag{70}$$
$$\tag{71}$$

We can instead control the flow on $\cos s\theta$:

$$\frac{d}{dt}\cos s\theta = -\dot{\theta} s \sin s\theta = \|P_A h^*\|^2 s r^{s-2}\sin^2 s\theta \tag{72}$$

and calculate the flow on v:

$$\dot{v} = 2\,\mathrm{Re}\{\langle A\dot{w}, Aw\rangle\} - 2r\dot{r} \tag{73}$$
$$= 2\left(r^s\cos s\theta - (v + r^2)^s - \|P_A h^*\|^2 r^s\cos s\theta + (v + r^2)^{s-1} r^2\right) \tag{74}$$
$$= 2(1 - \|P_A h^*\|^2) r^s\cos s\theta - 2(v + r^2)^{s-1} v \tag{75}$$

Finally, introducing the notation $\delta = 1 - \|P_A h^*\|^2$, we have

$$\dot{r} = (1-\delta) r^{s-1}\cos s\theta - (v + r^2)^{s-1} r \tag{76}$$
$$\frac{d}{dt}\cos s\theta = (1-\delta) s r^{s-2}(1 - \cos^2 s\theta) \tag{77}$$
$$\dot{v} = 2\delta r^s\cos s\theta - 2(v + r^2)^{s-1} v \tag{78}$$

E PROOF OF THEOREM 4.2

We will use the following facts repeatedly in the below arguments.

First, because $\dot{r} \geq 0$ when $r = 0$, and $\dot{r} \leq 0$ when $r = 1$, it follows that r can never leave the range $[0, 1]$. Furthermore, note that $\cos s\theta$ is always non-decreasing.

E.1 CASE $s = 1$

In the setting with information complexity equal to 1, we immediately have the following identities:

$$\dot{r} = (1 - \delta)\cos\theta - r \tag{79}$$

$$\frac{d}{dt}\cos\theta \geq (1 - \delta)(1 - \cos^2\theta) \tag{80}$$

$$\dot{v} \leq 2\delta - 2v \tag{81}$$

Let us address v first. From our assumptions, $\delta < \epsilon$, and so when $v \geq \epsilon$, $\dot{v}$ is negative. It follows that a trajectory that begins below ϵ cannot ever exceed ϵ. In other words, if $v_0 \leq \epsilon$, v can never exceed ϵ and we've achieved optimality.

Otherwise, supposing $v_0 > \epsilon$, consider values of t where $v_t > \delta$ so that the RHS of the inequality of $\dot{v}$ is strictly negative and we may write:

$$\frac{\dot{v}}{\delta - v} \geq 2 \tag{82}$$

Integrating from 0 to t gives that

$$-\log|\delta - v_t| - (-\log|\delta - v_0|) \geq 2t \tag{83}$$

which yields the bound

$$v_t \leq \delta + (v_0 - \delta)e^{-2t} \leq \delta + e^{-2t} \tag{84}$$

By Lemma H.1,

$$\cos\theta_t \geq \tanh((1 - \delta)t) \tag{85}$$

Finally, we consider r.

Choose $T_1 = \inf\{t \geq 0 : v_t \leq \epsilon, \cos\theta_t \geq \frac{1-\epsilon/2}{1-\delta}\}$, and $T_2 = \inf\{t \geq T_1 : r_t \geq 1 - \epsilon\}$. Note that one can easily confirm that $T_1 \leq O\left(\log\frac{1}{\epsilon}\right)$

Then for all $t \in [T_1, T_2)$, we have

$$\dot{r}_t = (1 - \delta)\cos\theta_t - r_t \geq 1 - \epsilon/2 - r_t \tag{86}$$

and the RHS is always non-negative.

Dividing by the RHS and integrating from T_1 to t gives

$$-\log(1 - \epsilon/2 - r_t) + \log(1 - \epsilon/2 - r_{T_1}) \geq t - T_1 \tag{87}$$

Rearranging gives

$$r_t \geq 1 - \epsilon/2 - (1 - \epsilon/2 - r_{T_1})e^{T_1 - t} \tag{88}$$

Note that by definition of T_2, it follows that

$$1 - r_t \leq \epsilon/2 + e^{T_1 - t} \tag{89}$$

So it follows that $T_2 \leq T_1 + \log \frac{2}{\epsilon}$.

Altogether, the total time to achieve ϵ optimality for all three variables is $O\left(\log \frac{1}{\epsilon}\right)$.

E.2 Case $s > 1$

In this case, because we cannot straightforwardly solve or bound the system of ODEs, we need to control rates in stages. We have a stopping time for one variable at a time, and use local monotonicity to ensure bounds on the remaining variables.

First Phase In the first stage, we consider the duration of time $T_1 = \inf\{t \geq 0 : v_t \leq v^*\}$ where $v^* := 2^{-s}6^{-2}s^{-2}r_0^4$, and bound the behavior of each variable. Below, we will consider $t \in [0, T_1]$.

To control the behavior or r, we consider the following manipulations:

$$\frac{d}{dt}\log r^2 = 2(1-\delta)r^{s-2}\cos s\theta - 2(v + r^2)^{s-1} \tag{90}$$

$$\frac{d}{dt}\log v = 2\delta\frac{r^s \cos s\theta}{v} - 2(v + r^2)^{s-1} \tag{91}$$

This implies

$$\frac{d}{dt}\log\frac{r^2}{v} = 2r^{s-2}\cos s\theta\left(1 - \delta - \delta\frac{r^2}{v}\right) \tag{92}$$

By definition, in this range of t we have $v_t > \frac{\delta}{1-\delta}$, so it follows that the RHS of this equation is always positive. Hence it follows that $\log \frac{r^2}{v}$ is increasing, and by monotonicity of $\log$, we have

$$\frac{r^2}{v} \geq \frac{r_0^2}{v_0} \geq r_0^2 \tag{93}$$

This implies that

$$\dot{r} = (1-\delta)r^{s-1}\cos s\theta - (v + r^2)^{s-1}r \tag{94}$$

$$\geq (1-\delta)r^{s-1}\cos s\theta - \left(\frac{r^2}{r_0^2} + r^2\right)^{s-1} r \tag{95}$$

$$\geq r^{s-1}\left((1-\delta)\cos s\theta - \left(\frac{1}{r_0^2} + 1\right)^{s-1} r^s\right) \tag{96}$$

Suppose it is true that $r \leq \frac{1}{6}r_0^2$, then it follows that:

$$r \leq \frac{r_0^2(1-\delta)\cos s\theta_0}{2} \tag{97}$$
$$\leq \frac{r_0^2(1-\delta)\cos s\theta_0}{r_0^2+1} \tag{98}$$
$$= \frac{(1-\delta)\cos s\theta_0}{\frac{1}{r_0^2}+1} \tag{99}$$
$$\leq \frac{((1-\delta)\cos s\theta)^{1/s}}{\left(\frac{1}{r_0^2}+1\right)^{\frac{s-1}{s}}} \tag{100}$$

So it follows that $\dot{r}$ will be positive whenever $r \leq \frac{1}{6}r_0^2$. We have $r_0 \geq \frac{1}{6}r_0^2$, it follows that $r_t \geq \frac{1}{6}r_0^2$ for $t \leq T_1$.

Finally we can control v by observing that, for $t \in [0, T_1]$, $v \geq v^* \geq (2\delta)^{1/s}$. Hence,

$$\dot{v} \leq 2\delta - 2v^s \leq -v^s \tag{101}$$

which implies

$$-\frac{\dot{v}}{v^s} \geq 1 \tag{102}$$

And integrating from 0 to $t \leq T_1$ gives

$$v_t^{-(s-1)} \geq \frac{1}{s-1}v_t^{-(s-1)} - \frac{1}{s-1}v_0^{-(s-1)} \geq t \tag{103}$$

Rearranging gives

$$v_t \leq t^{-\frac{1}{s-1}} \tag{104}$$

This gives a bound on $T_1 \leq (v^*)^{-(s-1)} = O(2^{s^2}r_0^{-4s})$

Lastly by monotonicity we have $\cos s\theta_{T_1} \geq \cos s\theta_0$.

So to summarize:

$$r_{T_1} \geq \frac{1}{6}r_0^2 \tag{105}$$
$$\cos s\theta_{T_1} \geq \cos s\theta_0 \tag{106}$$
$$v_{T_1} \leq v^* \tag{107}$$

Furthermore, we've actually proven that $v_t \leq v^*$ for all $t \geq T_1$, which we will use in subsequent phases.

Second Phase We define $T_2 = \inf\{t \geq T_1 : r_t \geq 1/5\}$. As before, if $r_{T_1} \geq 1/5$ then $T_2 = 0$ and we can skip to the next phase, so we assume otherwise.

Using the identity $(1+x)^k \leq 1 + 2^k x$ which holds for any $x \in [0, 1]$ and $k \geq 1$, observe that the ODE governing r can now be bounded as:

$$\dot{r} = (1-\delta)\cos s\theta r^{s-1} - (v+r^2)^{s-1} r \tag{108}$$

$$\geq (1-\delta)\cos s\theta_0 r^{s-1} - \left(\frac{v}{r^2}+1\right)^{s-1} r^{2s-1} \tag{109}$$

$$\geq (1-\delta)\cos s\theta_0 r^{s-1} - \left(1+2^{s-1}\frac{v}{r^2}\right) r^{2s-1} \tag{110}$$

$$\geq \frac{1-\delta}{2} r^{s-1} - \left(1+\frac{r_0^4}{2s^2(6r)^2}\right) r^{2s-1} \tag{111}$$

where in the last step we use that $v \leq v^*$ and plug in the definition of v^* and the bound $\cos s\theta_0 \geq 1/2$.

Consider any t when $r = \frac{1}{6}r_0^2$, and observe that the above inequality implies $\dot{r} > 0$. Because $r_{T_1} \geq \frac{1}{6}r_0^2$, this implies we will always have $r \geq \frac{1}{6}r_0^2$ for larger values of t, and we may bound:

$$\dot{r} \geq \frac{1-\delta}{2} r^{s-1} - \left(1+\frac{1}{2s^2}\right) r^{2s-1} \tag{112}$$

Hence, we can apply Lemma H.2 with $a = (1-\delta)/2$, $b = 1 + \frac{1}{2s^2}$, where $k^2 = (a/b)^2 \geq 1/5$, and using the initialization of r_{T_1}. This grants the bound that $T_2 \leq T_1 + O(s^4 r_{T_1}^{-s+1}) = T_1 + O(6^s r_0^{-2s+2})$.

Therefore the new summary is:

$$r_{T_2} \geq 1/5 \tag{113}$$

$$\cos s\theta_{T_2} \geq \cos s\theta_0 \tag{114}$$

$$v_{T_2} \leq v^* \tag{115}$$

Third Phase We define $T_3 = \inf\{t \geq T_2 : \cos s\theta_t \geq \frac{1-\frac{1}{4s^4}}{1-\delta}\}$

First of all, note that the bound on r derived in the last phase required lower bounding $\cos s\theta$ by $\cos s\theta_0$. Since $\cos s\theta$ is non-decreasing, that bound is still true by an identical argument.

So we can bound the ODE for θ:

$$\frac{d}{dt}\cos s\theta = (1-\delta)sr^{s-2}(1-\cos^2 s\theta) \tag{116}$$

$$\geq (1-\delta)s(1/5)^{s-2}(1-\cos^2 s\theta) \tag{117}$$

Note that by lemma H.1 with $k = (1-\delta)s(1/5)^{s-2}$, we have

$$T_3 \leq T_2 + O(5^s \log s) \tag{118}$$

The bound $v \leq v^*$ continues to hold. In summary, we now have:

$$r_{T_3} \geq 1/5 \tag{119}$$

$$\cos s\theta_{T_3} \geq \frac{1-\frac{1}{4s^4}}{1-\delta} \tag{120}$$

$$v_{T_3} \leq v^* \tag{121}$$

Fourth Phase We define $T_4 = \inf\{t \geq T_3 : r_t \geq r^*\}$ where $r^* := 1 - \frac{1}{s^2}$. Again, consider the non-trivial case where $T_4 \neq 0$.

Because the bound on v is the same, and the bound on $\cos s\theta$ is better than before, we can now bound the ODE of r similarly to the second phase:

$$\dot{r} \geq \left(1 - \frac{1}{4s^4}\right) r^{s-1} - \left(1 + \frac{1}{2s^2}\right) r^{2s-1} \tag{122}$$

Applying Lemma H.2 with $k = \frac{1-\frac{1}{4s^4}}{1+\frac{1}{2s^2}} = 1 - \frac{1}{2s^2}$, we have:

$$T = \inf\{t \geq T_3 : r \geq k^2\} \leq T_3 + O(5^s \log s) \tag{123}$$

Finally, note that $k^2 = \left(1 - \frac{1}{2s^2}\right)^2 \geq 1 - \frac{1}{s^2}$, which implies that $T_4 \leq T$.

Thus we have:

$$r_{T_4} \geq r^* \tag{124}$$

$$\cos s\theta_{T_4} \geq \frac{1 - \frac{1}{4s^4}}{1 - \delta} \tag{125}$$

$$v_{T_4} \leq v^* \tag{126}$$

Fifth Phase We define $T_5 = \inf\{t \geq T_4 : \cos s\theta_t \geq \frac{1-\epsilon/2}{1-\delta}, v_t \leq v^\dagger\}$ where $v^\dagger = 2^{-s}(\epsilon/2)(r^*)^2$.

Again, since $\cos s\theta$ is increasing and v is always less than v^*, the bound on $r \geq r^*$ established in the last step will stay true.

Thus, by the identity $r^k \geq (r^*)^k = \left(1 - \frac{1}{s^2}\right)^k \geq 1 - \frac{k}{s^2}$ we have the ODE inequalities:

$$\frac{d}{dt} \cos s\theta = (1 - \delta) s r^{s-2} (1 - \cos^2 s\theta) \tag{127}$$

$$\geq (1 - \delta) s \left(1 - \frac{1}{s}\right) (1 - \cos^2 s\theta) \tag{128}$$

$$\dot{v} = 2\delta r^s \cos s\theta - 2(v + r^2)^{s-1} v \tag{129}$$

$$\leq 2\delta - 2\left(1 - \frac{2(s-1)}{s^2}\right) v \tag{130}$$

It is easy to see that we'll have the bound

$$T_5 \leq T_4 + O\left(\log \frac{1}{\epsilon}\right) \tag{131}$$

and in summary

$$r_{T_5} \geq r^* \tag{132}$$

$$\cos s\theta_{T_5} \geq \frac{1 - \epsilon/2}{1 - \delta} \tag{133}$$

$$v_{T_5} \leq v^\dagger \tag{134}$$

Sixth Phase We define $T_6 = \inf\{t \geq T_5 : r_t \geq 1 - \epsilon\}$, and assume the non-trivial setting where $T_6 \neq 0$.

Note that $\dot{v}$ is negative when $v = v^\dagger$, so the bound $v \leq v^\dagger$ remains true for $t \geq T_5$. Thus, we can control the ODE of r one more time:

$$\dot{r} = (1-\delta)r^{s-1}\cos s\theta - (v+r^2)^{s-1}r \tag{135}$$

$$\geq (1-\delta)r^{s-1}(1-\epsilon/2) - \left(1+\frac{v}{r^2}\right)^{s-1} r \tag{136}$$

$$\geq (1-\epsilon/2)r^{s-1} - \left(1+2^s\frac{v^\dagger}{r^2}\right) r^{2s-1} \tag{137}$$

$$\geq (1-\epsilon/2)r^{s-1} - \left(1+\epsilon/2\frac{(r^*)^2}{r^2}\right) r^{2s-1} \tag{138}$$

One can confirm that when $r = r^*$, the RHS of the above inequality is positive, so $\dot{r} \geq 0$. Thus, since $r_{T_5} \geq r^*$, it will always be the case that $r \geq r^*$ for $t \geq T_5$, so as before we bound:

$$\dot{r} \geq (1-\epsilon/2)r^{s-1} - (1+\epsilon/2)r^{2s-1} \tag{139}$$

By Lemma H.2, we have that

$$T_6 \leq T_5 + O\left(\log\frac{1}{\epsilon}\right) \tag{140}$$

and thus we've achieved ϵ optimality for all three of our variables.

F Proof of Lemma 4.5

Remind from Proposition 2.2 that $A \in \mathbb{C}^{\infty\times M}$ is of the form

$$A_{km} = c_k\sqrt{k}a_m^k \tag{141}$$

where we assume $c_k > 0$, and $a_m \sim S^1$. Note that

$$1 - \|P_A h^*\|^2 = \|P_A^\perp h^*\|^2 \tag{142}$$

$$= \min_w \|Aw - h^*\|^2 \tag{143}$$

$$\tag{144}$$

so we need to choose a candidate value of w.

Consider the block decomposition

$$A = \left[\frac{B}{C}\right] \tag{145}$$

where $B \in \mathbb{C}^{N\times M}$ and $C \in \mathbb{C}^{\infty\times M}$. Suppose we decompose $h^* = \left[\frac{u}{0}\right]$ where $u \in \mathbb{C}^N$. Then if we apply the pseudoinverse and define $w = B^+u$, observe:

$$Aw = \left[\frac{B}{C}\right] B^+u \tag{146}$$

$$= \left[\frac{BB^+u}{CB^+u}\right] \tag{147}$$

Observe that we can decompose $B = DX$ where D is a diagonal matrix such that $D_{kk} = c_k\sqrt{k}$ and $X_{km} = a_m^k$. Since $N < M$, one can see X is a rectangular Vandermonde matrix evaluated on $\{a_m\}_{m=1}^M$. Almost surely, these values are all pairwise distinct, which implies that X has linearly independent rows. Since D is diagonal with no zeros along the diagonal, B also has linearly independent rows. This condition implies $BB^+ = I$. So we have

$$Aw = \begin{bmatrix} u \\ \hline CB^+u \end{bmatrix} \tag{148}$$

Remember $\|u\| = \|h^*\| = 1$, as u is the first N elements of h^* and hence still only supported on the first $\sqrt{N}$ elements. Because $B^+ = X^+D^{-1}$, we have:

$$\|CB^+u\| \leq \|C\|\|X^+\|\|D^{-1}u\| \tag{149}$$
$$\tag{150}$$

We can now go about bounding these norms.

Since u is only supported on the first $\sqrt{N}$ elements and $\|u\| = 1$, it follows $\|D^{-1}u\| \leq \max_{1\leq k\leq\sqrt{N}} \left|\frac{1}{c_k\sqrt{k}}\right| = \frac{1}{\sigma_-}$.

By Lemma H.4, we have the bound

$$\|X^+\| \leq O\left(\frac{1}{\sqrt{M}}\right) \tag{151}$$

Finally for any $\hat{w} \in \mathbb{C}^M$ with $\|\hat{w}\| = 1$, we have by Cauchy-Schwarz:

$$\|Cw\|^2 = \sum_{k=N+1}^{\infty} \left|\sum_{m=1}^{M} \hat{w}_m c_k\sqrt{k} a_m^k\right|^2 \tag{152}$$
$$\leq \sum_{k=N+1}^{\infty} \|\hat{w}\|^2 \sum_{m=1}^{M} \left|c_k\sqrt{k}\right|^2 \tag{153}$$
$$= M \sum_{k=N+1}^{\infty} k|c_k|^2 \tag{154}$$
$$\leq Me^{-\Omega(\sqrt{N})} \tag{155}$$

where we use in the last step Assumption 4.4.

With these bounds, we clearly have

$$1 - \|P_A h^*\| \leq \|Aw - h^*\|^2 \tag{156}$$
$$= \left\| \begin{bmatrix} u \\ \hline CB^+u \end{bmatrix} - \begin{bmatrix} u \\ \hline 0 \end{bmatrix} \right\|^2 \tag{157}$$
$$\leq \|CB^+u\|^2 \tag{158}$$
$$\leq \frac{M}{\sqrt{M}\sigma_-} e^{-\Omega(\sqrt{N})} \tag{159}$$

Because $M = O(N^3)$, and we've assumed $1/\sigma_-$ is polynomial in N, this bound can be written as $e^{-\Omega(\sqrt{N})}$ for possibly different constants in the big O notation.

G PROOF OF LEMMA 4.6

Remind that $m_0 = \langle Aw_0, h^* \rangle = \frac{1}{\|Aw\|}\langle Aw, h^* \rangle$. Because the complex Gaussian is invariant to multiplication by an unit modulus complex number, it follows that θ_0 is independent of r_0 and uniformly distributed on S^1. Because s is a positive integer, $s\theta_0$ is also uniformly distributed on S^1, and hence $P(\cos s\theta_0 \geq 1/2) = 1/3$. And by our choice of normalization, $v_0 = 1 - r_0^2$ automatically. So it only remains to prove the first statement is true with high probability.

We remind that $r_0 = \frac{|\langle Aw, h^* \rangle|}{\|Aw\|}$. By Cauchy-Schwartz, it's clear that $r_0 \leq 1$, so only the lower bound is non-trivial. If we use the same notation to decompose the matrix A as in the proof of Lemma 4.5, it's clear that

$$|\langle Aw, h^* \rangle| = |\langle Bw, u \rangle| \tag{160}$$
$$= |\langle w, B^\dagger u \rangle| \tag{161}$$

If we condition on B, then by rotation invariance of the Gaussian, note that $|\langle w, B^\dagger u \rangle|$ is distributed identically to $|g|\|B^\dagger u\|$ where g is sampled from a one dimensional complex Gaussian.

By the argument in Lemma 4.6, since u is only supported on the first $\sqrt{N}$ elements, note that:

$$\|B^\dagger u\| = \|X^\dagger D^\dagger u\| \tag{162}$$
$$\geq \sigma_N(X)\|D^\dagger u\| \tag{163}$$
$$\geq \sigma_N(X)\sigma_- \tag{164}$$
$$\geq \sigma_- O(\sqrt{M}) \tag{165}$$

with probability $1 - 2\exp(-\Omega(N))$ by Lemma H.4

Lastly, we need to control

$$\|Aw\| \leq \|Bw\| + \|Cw\| \leq (\|B\| + \|C\|)\|w\| \tag{166}$$

And we can write again by Lemma H.4, with similarly high probability:

$$\|B\| = \|DX\| \tag{167}$$
$$\leq \|D\|\|X\| \tag{168}$$
$$\leq \sigma_+\sigma_1(X) \tag{169}$$
$$\leq \sigma_+ O(\sqrt{M}) \tag{170}$$

Combining this with the bound on $\|C\|$ we derived in Lemma 4.6, and the concentration on $\|w\|$ from Lemma H.3 we have with probability $1 - 2\exp(-\Omega(N))$:

$$\|Aw\| \leq \left(\sigma_+ O(\sqrt{M}) + e^{-\Omega(\sqrt{N})}\right) O(\sqrt{M}) \tag{171}$$

Finally we can say that with probability $1 - 2\exp(-\Omega(N))$

$$r_0 \geq c\frac{\sigma_-}{\sigma_+\sqrt{M}} \tag{172}$$

for some universal constant c.

H AUXILIARY LEMMAS

H.1 DYNAMICS INEQUALITY LEMMAS

The following lemmas provide bounds on our dynamics that we can apply multiple times in different phases of the proof. Both of these lemmas are essentially special cases of the Bihari-LaSalle Inequality (Bihari, 1956), but because the proofs are much simplified due to our setting, and for completeness, we include the proofs below.

Lemma H.1. *Consider θ with the differential inequality*

$$\frac{d}{dt}\cos s\theta \geq k(1-\cos^2 s\theta) \tag{173}$$

with $\cos s\theta_0 \geq 1/2$. Then we have

$$\cos s\theta_t \geq \tanh(kt) \tag{174}$$

and hence if $T = \inf\{t \geq 0 : \cos s\theta_t \geq c\}$, then $T \leq \frac{1}{2k}\log\frac{2}{1-c}$.

Proof. Clearly the RHS of the inequality is always positive, so we may write:

$$\frac{\frac{d}{dt}\cos s\theta}{1-\cos^2 s\theta} \geq k \tag{175}$$

and integrating from 0 to t gives

$$\tanh^{-1}(\cos s\theta_t) - \tanh^{-1}(\cos s\theta_0) \geq kt \tag{176}$$

Note $\tanh^{-1}(\cos s\theta_0) \geq 0$, so $\cos s\theta_t \geq \tanh(kt)$. Since $\cos s\theta_t$ is increasing, it follows that

$$T \leq \frac{\tanh^{-1}(c)}{k} \tag{177}$$

And using the closed form of $\tanh^{-1}$(c) for $|c| < 1$ implies

$$T \leq \frac{1}{2k}\log\frac{1+c}{1-c} \tag{178}$$

$$\leq \frac{1}{2k}\log\frac{2}{1-c} \tag{179}$$

□

Lemma H.2. *Consider $s \geq 2$. Suppose we have constants $0 < a < b$ and a function r of time t with differential identity:*

$$\dot{r} \geq ar^{s-1} - br^{2s-1} \tag{180}$$

Furthermore, assume $0 < r_0$ and it always the case that $r \leq 1$.

Let $k = \frac{a}{b}$, and $T = \inf\{t \geq 0 : r \geq k^2\}$, then:

$$T \leq \frac{1}{bk^2}\left(\frac{2k}{r_0^{s-1}} + \log\frac{1}{1-k}\right) \tag{181}$$

Proof. If $r_0 \geq k^2$, then $T = 0$ and the bound is obviously true. So assume $r_0 < k^2 \leq k^{\frac{1}{s-1}}$, where the second inequality follows from the facts that $k < 1$ and $s \geq 2$.

Consider the change of variables $y = r^{s-1}/k$:

$$\dot{y} = \frac{1}{k}(s-1)r^{s-2}\dot{r} \tag{182}$$

$$\geq \frac{1}{k}(s-1)(ar^{2s-3} - br^{3s-3}) \tag{183}$$

$$\geq \frac{1}{k}(ar^{2s-2} - br^{3s-3}) \tag{184}$$

$$= \frac{b}{k}(kr^{2s-2} - r^{3s-3}) \tag{185}$$

$$= \frac{b}{k}(k^3y^2 - k^3y^3) \tag{186}$$

$$= bk^2y^2(1-y) \tag{187}$$

For $t \in [0, T)$, the RHS will always be positive, so we can write

$$\frac{\dot{y}}{y^2(1-y)} \geq bk^2 \tag{188}$$

Simple algebra lets us rewrite:

$$\frac{\dot{y}}{y} + \frac{\dot{y}}{y^2} + \frac{\dot{y}}{1-y} \geq bk^2 \tag{189}$$

And integrating from 0 to t gives

$$\log y_t - \log y_0 - \frac{1}{y_t} + \frac{1}{y_0} - \log(1-y_t) + \log(1-y_0) \geq bk^2t \tag{190}$$

Remind that $\frac{1}{y_t} > 0$ and collecting terms, we have:

$$-\log\left(\frac{1}{y_t} - 1\right) + \log\left(\frac{1}{y_0} - 1\right) \geq bk^2t - \frac{1}{y_0} \tag{191}$$

Taking exponentials and simple bounds:

$$\frac{1}{y_t} - 1 \leq \frac{1}{y_0}\exp\left(-bk^2t + \frac{1}{y_0}\right) \tag{192}$$

Rearranging and reminding $y_t = r_t^{s-1}/k$

$$\frac{k}{1 + \frac{1}{y_0}\exp\left(-bk^2t + \frac{1}{y_0}\right)} \leq r_t^{s-1} \leq r_t \tag{193}$$

To finish the proof, we'll show that $r_t \geq k^2$ is implied by a condition on t. Suppose that

$$t \geq \frac{1}{bk^2}\left(\frac{2}{y_0} + \log\frac{1}{1-k}\right) \tag{194}$$

Then using the fact that $k < 1$, and $\log x < x$ for all $x > 0$, it follows

$$t \geq \frac{1}{bk^2}\left(\frac{1}{y_0} + \log\frac{1}{y_0} + \log\frac{k}{1-k}\right) \tag{195}$$

$$\geq \frac{1}{bk^2}\left(\frac{1}{y_0} + \log\frac{1}{y_0\left(\frac{1}{k}-1\right)}\right) \tag{196}$$

Rearranging implies that

$$k \leq \frac{1}{1 + \frac{1}{y_0}\exp\left(-bk^2t + \frac{1}{y_0}\right)} \tag{197}$$

and plugging this into Equation 193 implies that $r_t \geq k^2$. Hence, the stopping time T obeys:

$$T \leq \frac{1}{bk^2}\left(\frac{2}{y_0} + \log\frac{1}{1-k}\right) \tag{198}$$

Plugging in the definition of y_0 gives the bound. □

H.2 Concentration Inequality Lemmas

We require a few very standard lemmas, adapting concentration inequalities to the complex setting.

Lemma H.3. *If w is drawn from the standard complex Gaussian on M dimensions, then*

$$P(|\|w\| - \sqrt{M}| \geq t) \leq 2\exp\left(-ct^2\right) \tag{199}$$

for some universal constant c.

Proof. Note that an equivalent way of sampling a complex Gaussian is $w = \frac{1}{\sqrt{2}}(w_R + iw_C)$ with w_R, w_C both sampled iid from a standard real Gaussian on M variables. Therefore

$$\|w\|^2 = \frac{\|w_R\|^2 + \|w_C\|^2}{2} \tag{200}$$

$$= \frac{1}{2}\left\|\begin{bmatrix} w_R \\ w_C \end{bmatrix}\right\|^2 \tag{201}$$

Note that $\hat{w} := \begin{bmatrix} w_R \\ w_C \end{bmatrix}$ is simply a standard Gaussian on $2M$ variables, so from Theorem 3.1.1 in Vershynin (2018):

$$P(|\|w\| - \sqrt{M}| \geq t) = P(|\|\hat{w}\| - \sqrt{2M}| \geq t\sqrt{2}) \tag{202}$$

$$\leq 2\exp\left(-ct^2\right) \tag{203}$$

for some universal constant c.

□

Lemma H.4. *Let $a_m \sim S^1$ be sampled iid, for $m = 1, \dots, M$, and define $X \in \mathbb{C}^{N \times M}$ as $X_{nm} = a_m^n$. Then if we choose $M = O(N^3)$, with probability $1 - 2\exp(-\Omega(N))$:*

$$\sigma_1(X) = \Theta(\sqrt{M}), \sigma_N(X) = \Theta(\sqrt{M}) \tag{204}$$

Proof. Note that the columns of X are independent, mean zero, and isotropic. Let X_m be the mth column, and consider any $v \in \mathbb{C}^M$ with $\|v\| = 1$. Note that $\|X_m\| = \sqrt{N}$, so it follows that

$$\|\langle X_m, v\rangle\|_{\psi_2} \leq \sqrt{N} \tag{205}$$

where $\|\cdot\|_{\psi_2}$ denotes the subgaussian norm (Vershynin, 2018). Hence, we can apply Theorem 4.6.1 from Vershynin (2018) to X^T. Note, although this proof assumes real-valued variables, the same arguments follow through with no change to complex variables given the subgaussian bound on $\|\langle X_m, v\rangle\|_{\psi_2}$. Hence,

$$\sqrt{M} - cN(\sqrt{N} + t) \leq \sigma_N(X) \leq \sigma_1(X) \leq \sqrt{M} + CN(\sqrt{N} + t) \tag{206}$$

for universal constants c, C and with probability $1 - 2\exp\left(-t^2\right)$. Choosing $t = \sqrt{N}$ and $M = O(N^3)$ gives the result.

□

H.3 Valid Activations

We quickly note a one simple choice of many possible activation functions that meets our criteria in Assumption 4.4.

Proposition H.5. *Let $\sigma(z) = \arctan \xi z + \xi z \arctan \xi z$ for $\xi = 1 - \frac{1}{\sqrt{N}}$. Then this activation satisfies Assumption 4.4, and $\sigma_+ \leq \sqrt{2}$, $\sigma_- \geq O\left(\frac{1}{N^{1/4}}\right)$.*

Proof. Observe that σ is analytic on the unit disk following from properties of $\arctan$, with the Laurent series

$$\sigma(z) = \xi z + \xi^2 z^2 - \frac{\xi^3}{3} z^3 - \frac{\xi^4}{3} z^4 + \dots \tag{207}$$

So the only coefficient equal to zero is the constant term. Moreover, if split into sequence of odd degree and even degree coefficients, both sequences are decreasing in absolute value, so we can instantly say that $\sigma_+ \leq \sqrt{2}$ and

$$\sigma_- = \min_{1 \leq k \leq \sqrt{N}} |c_k| \sqrt{k} \tag{208}$$

$$\geq \frac{\left(1 - \frac{1}{\sqrt{N}}\right)^{\sqrt{N}}}{\sqrt{N}} N^{1/4} = O\left(\frac{1}{N^{1/4}}\right) \tag{209}$$

Moreover, we can calculate:

$$\sum_{k=N+1}^{\infty} k|c_k|^2 \leq \sum_{k=N+1}^{\infty} \frac{k\xi^{k-1}}{(k-1)^2} \tag{210}$$

$$\leq \sum_{k=N+1}^{\infty} \xi^{k-1} \tag{211}$$

$$\leq \frac{\xi^N}{1 - \xi} \tag{212}$$

$$\leq e^{-\Omega(\sqrt{N})} \tag{213}$$

□