# Tensor Programs III: Neural Matrix Laws

**Greg Yang**
Microsoft Research AI
gregyang@microsoft.com

## Abstract

In a neural network (NN), *weight matrices* linearly transform inputs into *preactivations* that are then transformed nonlinearly into *activations*. A typical NN interleaves multitudes of such linear and nonlinear transforms to express complex functions. Thus, the (pre-)activations depend on the weights in an intricate manner. We show that, surprisingly, (pre-)activations of a randomly initialized NN become *independent* from the weights as the NN's widths tend to infinity, in the sense of *asymptotic freeness* in random matrix theory. We call this the *Free Independence Principle (FIP)*, which has these consequences: 1) It rigorously justifies the calculation of asymptotic Jacobian singular value distribution of an NN in Pennington et al. [36, 37], essential for training ultra-deep NNs [48]. 2) It gives a new justification of *gradient independence assumption* used for calculating the *Neural Tangent Kernel* of a neural network. FIP and these results hold for any neural architecture. We show FIP by proving a Master Theorem for any Tensor Program, as introduced in Yang [50, 51], generalizing the Master Theorems proved in those works. As warmup demonstrations of this new Master Theorem, we give new proofs of the semicircle and Marchenko-Pastur laws, which benchmarks our framework against these fundamental mathematical results.

## 1 Introduction

A neural network (NN), at a high level, is an interleaved composition of linear and nonlinear transformations. Its *weights* are the matrices involved in the linear transformations. The image (resp. input) of nonlinear transformations are called *activations* (resp. *preactivations*). Thus, *a priori*, (pre-)activations depend in complicated ways on the weights. We show that, surprisingly, the (pre-)activations become roughly *independent* from the weights in the sense of random matrix theory, when the NN is randomly initialized and as its width tends to infinity. More formally, we prove that the weights are *asymptotically free* from the diagonal matrices whose diagonals are the images of preactivation vectors under any bounded coordinatewise function.[1] This result holds for any neural architecture (i.e. architectural universality)[2] . We call this result the *Free Independence Principle* (FIP). A major application of FIP is in rigorously justifying a prevalent free independence assumption in the calculation of NN Jacobian singular value distribution [25, 36, 37, 43, 48], as we overview below.

Besides FIP, we also prove several other interesting results. In the next few subsections, we discuss the context required to state them. A summary of all of our contributions appears at the end of this introduction, and the reader in a hurry may feel free to skip directly to it.

[1]Note, however, that we make no claim about trained weights, just random weights.

[2]In this work, *architecture* refers to the network topology along with the random initialization of parameters, as specified by a NETSOR$\top$ program (Definition 2.1).

### 1.1 Random Matrix Theory in Deep Learning

Random matrix theory (RMT) has a successful history of being applied in deep learning [25, 36, 37, 43, 48]. For example, RMT has been used to calculate the Jacobian[3] singular value distribution of a wide neural network, which is an important indicator of its architectural soundness: If this distribution becomes very diffuse as the network gets deeper (i.e. more layers), then an error signal (i.e. the gradient) will be badly distorted as it is backpropagated. On the other hand, if the distribution concentrates around 1 even when the network gets deeper, then the error signal is largely preserved, and all layers of the networks receive adequate signal to improve. This idea is known as *dynamical isometry* [36] and has been successfully applied in practice to train ultra-deep networks (as deep as 10,000 layers! [48]).

This calculation of the NN Jacobian singular values distribution can be seen as a vast nonlinear generalization of the classical problem of calculating the singular value distribution for an iid random matrix. The solution of this latter problem is famously given by (the square root of) the Marchenko-Pastur distribution. A related classical random matrix law is the semicircle law, which says that the eigenvalue distribution of the sum $A + A^\top$ of a random real iid matrix $A$ and its transpose $A^\top$ is asympotically shaped like a semicircle. Like the utility of NN Jacobian singular values, Marchenko-Pastur and the semicircle law have both been impactful in science and engineering and are considered two fundamental laws of random matrix theory. In fact, the semicircle law was first motivated by a study of the distribution of energy levels of an atom [21].

Up until recently, the study of the NN Jacobian singular value distribution lacked a rigorous foundation. The first calculation of it assumed (free) independence (between the weights and the preactivations) that was only empirically checked but not proved. It was unknown how widespread this phenomenon actually was. One purpose of this work is filling in that hole: FIP reveals the architectural universality of this free independence (i.e. the neural network can be as complex as any modern manifestations and the principle still holds).

### 1.2 New Approach to Random Matrix Theory

Another purpose of this work is to illustrate a new way to go about random matrix theory, one that allows us to methodically attack the nonlinear problems of deep neural networks. In contrast, classical methods of random matrix theory typically heavily rely on the linearity of classical random matrix ensembles. The *expansion technique* is one such method. We review how it's used in linear problems and show it is ineffective for (nonlinear) neural networks problems.

**The expansion technique of classical RMT** For example, a typical quantity to calculate is the expectation of the trace of some matrix power $\mathrm{tr}(A^\top A)^k$ for some iid random matrix $A$. One common way to proceed is expanding this expression in terms of monomials of the entries of $A$, and then counting the different kinds of monomials that arise: such as monomials multilinear in the entries of $A$, or monomials where every entry appears quadratically, etc. Given this expansion, the trace expectation then follows easily.

**The expansion technique is ineffective in neural network problems** Now, in the (nonlinear) case of neural networks, the matrix $A$ is usually something more complicated. Take, for example, $A = D_2 B D_1 B$ for some iid matrix $B$ and diagonal matrices $D_i$ that depend nonlinearly on $B$. This dependence can, for example, take the following form: for some iid vector $v$ and $\phi = \tanh$ (to be applied coordinatewise), we define $u_1 = \phi(Bv), u_2 = \phi(Bu_1)$ and set $D_i$ to have diagonal $u_i$. Such ensemble $A$ commonly appears as the Jacobian of a neural network. Inspired by the classical expansion technique described above, one may attack this problem by expanding $\mathrm{tr}(A^\top A)^k$ in terms of entries of $A$ or in terms of entries of $B$. Because the entries of $A$ are correlated in complicated ways, it will be hard to use the expansion in $A$. On the other hand, we cannot even expand $\mathrm{tr}(A^\top A)^k$ cleanly in $B$ because of the nonlinear dependence of $D_i$ on $B$ (particularly because of $\phi$)[4].

[3]Here we are concerned with the input-output Jacobian of a neural network on a fix input. Some works consider the Jacobian with respect to the parameters, where the Jacobian has dimension #parameters $\times$ #data points (e.g. [35]). This is a related but distinct case from the input-output Jacobian, which is what we focus on in this work.

[4]We can Taylor expand $D_i$ in terms of entries of $B$ but that gets ugly really fast.

Thus the usual expansion technique runs into a wall very quickly.

**Our proposed method** We express the trace $\mathrm{tr}(A^\top A)^k$ as an expectation $\mathbb{E}_v\, v^\top (A^\top A)^k v$ where $v$ is a standard Gaussian vector[5]. Then we inductively *analyze* the vectors $Av, A^\top Av, AA^\top Av, (A^\top A)^2 v, \ldots, (A^\top A)^k v$, in a way we will soon describe below. Finally, this analysis of $(A^\top A)^k v$ will allow us to calculate $\mathbb{E}_v\, v^\top (A^\top A)^k v$.

It turns out that the vectors $v, Av, A^\top Av, AA^\top Av, (A^\top A)^2 v, \ldots, (A^\top A)^k v$ will

1. all have *approximately iid* coordinates in the large dimension limit, in a suitable sense, i.e. $\{(Av)_\alpha\}_\alpha$ are approximately iid, and similarly for $\{(A^\top Av)_\alpha\}_\alpha$, $\{((A^\top A)^2 v)_\alpha\}_\alpha$, etc, where $\alpha$ denotes coordinate index;
2. but the sequence of coordinates $v_\alpha, (Av)_\alpha, (A^\top Av)_\alpha, \ldots, ((A^\top A)^k v)_\alpha$ are correlated for each fixed index $\alpha$, in the same way across all $\alpha$.

By *analyze*, we mean to calculate such correlations. Then $\mathbb{E}_v\, v^\top (A^\top A)^k v$ is given by the correlation between $v_\alpha$ and $((A^\top A)^k v)_\alpha$.

**Example** Let us use the classical example of iid $A$ to make these two points more concrete.

If $v \in \mathbb{R}^n$ is a standard Gaussian vector, and $A \in \mathbb{R}^{n\times n}$, $A_{\alpha\beta} \sim \mathcal{N}(0, 1/n)$ is iid and independent from $v$, then it's not hard to see $Av$ has coordinates which tend to iid Gaussians via some kind of central limit argument (illustraing point (1)). Next, we will soon see intuitively that $A^\top Av$ is asymptotically the sum of $v$ and a random Gaussian vector with iid coordinates, independent from $v$ (illustrating point (1)). Thus $A^\top Av$ is *coordinatewise* correlated with $v$, illustrating point (2).

To get this intuition, write each coordinate

$$(A^\top Av)_\alpha = \sum_{\beta,\gamma} A_{\beta\alpha} A_{\beta\gamma} v_\gamma = \sum_\beta A_{\beta\alpha}^2 v_\alpha + \sum_{\beta;\gamma\neq\alpha} A_{\beta\alpha} A_{\beta\gamma} v_\gamma, \quad \beta, \gamma \in [n]. \tag{1}$$

Because each $A_{\beta\alpha}$ has variance $1/n$, the first sum will converge via law of large numbers to $v_\alpha$. In the second sum, one can note that each summand is uncorrelated with others (and higher order correlations drop off rapidly with $n$), so one may expect the second sum to converge to a Gaussian through some central limit behavior. We can calculate that the second sum $\sum_{\beta;\gamma\neq\alpha} A_{\beta\alpha} A_{\beta\gamma} v_\gamma$ for different $\alpha$s will be uncorrelated, so their limits for different $\alpha$ should be independent Gaussians. Likewise, it should be asymptotically independent from $v$ for the same reason.

This finishes our example and illustrates the overarching philosophy of our proposed method. But it is still tedious if we have to manually derive the correlation between $v_\alpha, (Av)_\alpha, (A^\top Av)_\alpha, \ldots, ((A^\top A)^k v)_\alpha$. In addition, it's not entirely clear at this point that this method can handle the nonlinear ensemble example $A = D_2 B D_1 B$ given above. This is where the Tensor Programs framework is crucial, which we overview now.

### 1.3 The Tensor Programs Framework

The Tensor Programs framework can be thought of as a clean, scalable, and rigorous packaging of the intuition explained in the last section. We first demonstrate how one can apply this framework to compute the correlations between $v_\alpha, (Av)_\alpha, (A^\top Av)_\alpha, \ldots, ((A^\top A)^k v)_\alpha$. Then we describe in more general terms what the framework is about.

**Calculating $\mathbb{E}_v\, v^\top (A^\top A)^k v$ with Tensor Programs** The framework associates a random variable $Z^u \in \mathbb{R}$ to each $u \in \{v, Av, A^\top Av, \ldots, (A^\top A)^k v\}$, representing the distribution of coordinates of $u$ in the large $n$ limit. It also comes with a set of symbolic rules to derive $Z^{Av}$ given $Z^v$, $Z^{A^\top Av}$ given $Z^{Av}$ and $Z^v$, and so on. For example, consistent with the intuition laid out in Section 1.2, $Z^v$ and $Z^{Av}$ are defined as independent standard Gaussians (since $v$ and $Av$ are asymptotically standard Gaussian vectors, independent from each other), and $Z^{A^\top Av} = Z^v + S$ for a standard Gaussian $S \in \mathbb{R}$ independent from $Z^v$ (and $Z^{Av}$). Finally, we can calculate $\lim_{n\to\infty} \mathbb{E}_v\, v^\top (A^\top A)^k v = \mathbb{E}\, Z^v Z^{(A^\top A)^k v}$.

[5] but this identity holds for any $v$ with iid, zero-mean, unit-variance entries

This set of $Z$-rules handle nonlinearities. For example, we have $Z^{\phi(v)} \stackrel{\text{def}}{=} \phi(Z^v)$ for any $\phi : \mathbb{R} \to \mathbb{R}$ (where $\phi$ on the LHS is applied coordinatewise to the vector $v$, but is applied in the usual sense to the scalar random variable $Z^v$ on the RHS.) This reflects the intuition that $Z^{\phi(v)}$ is the distribution of $\phi(v)$'s coordinates. As such, it is not hard to see our proposed method in Section 1.2 can deal with nonlinear random matrix ensembles like the example $A = D_2 B D_1 B$ given there. The full description of these rules can be found in Box 1, but we will not dive into more details here.

**Tensor Programs in general** While a major focus of this paper is on RMT, the original motivation of Tensor Programs is in understanding wide neural networks. There were 3 lines of research that molded the framework, which we briefly overview so we can later state some other contributions of this paper: *(Research Line 1)* In 1994, Radford Neal [31] discovered that shallow neural networks with random weights converge to Gaussian processes as their widths tend to infinity. This Neural Network-Gaussian Process correspondence has become a fascinating topic ever since, with a flurry of activities in recent years extending it [9, 17, 22, 23, 28, 32, 47]. *(Research Line 2)* Relatedly, a line of work, closely connected to the study of Jacobian singular values mentioned above, researched how signal propagates inside a neural network with random weights [7, 14–16, 36, 38–40, 52–54], and use such insights to improve the initialization scheme or architecture of neural networks. *(Research Line 3)* Recently, Jacot et al. and others [2–5, 11, 19, 24, 55] found that, when trained with gradient descent, wide, fully connected neural networks evolve like linear models. This resolved many open questions regarding the training and generalization of neural networks, and was followed by intense activity generalizing these results to more modern neural architectures [1, 5, 12, 18, 26, 49].

In all three lines of research, each new architecture required a new paper. For example, going from fully connected networks to convolutional neural networks required careful thinking about how the newly added dimension of pixel position in the latter changes the argument for the former. It was the expectation that every architecture is special in some way, and extending the theory to catch up with modern practice would require a paper covering every major architectural advancement starting from convolutional neural networks (e.g. normalization layers, attention, etc). Considering the breakneck pace of empirical deep learning, this "catch-up" might never happen.

It was against this backdrop that the Tensor Programs framework [49–51] surprisingly showed all three lines of research can be unified into one and simultaneously be generalized to all practically relevant neural architectures, now or in the future. There are two major insights that make this possible:

**Insight 1** Every computation done in deep learning can be written as a program (i.e. a composition) of matrix multiplication and coordinatewise nonlinearities[6] (for example, a simple neural network can be written this way as $h = Wx, y = \phi(h)$ for some activation function $\phi$, like $\tanh$, applied coordinatewise; the example of $A = D_2 B D_1 B$ given above can be written likewise).

**Insight 2** In every such computation, if the matrices are randomized, then every vector computed over the course of such a program has roughly iid coordinates when the matrices' sizes are large. However, for each $\alpha$, the $\alpha$th coordinates of different vectors will in general be correlated, in the same way across $\alpha$, and in a way that can be inductively calculated from the program structure. The machinery of Tensor Programs allows one to mechanically perform this calculation.

Note that insight 2 is a generalization of (1) and (2) in Section 1.2, and is reflected in the $\mathbb{E}_v\, v^\top (A^\top A)^k v$ example above. Given these two insights, it then becomes rather straightforward to generalize the three lines of research to any relevant architecture.

### 1.4 Tensor Programs Master Theorems

The Tensor Programs framework was first laid out in (the unpublished manuscript) Yang [49] but was written densely. Subsequently, [50] and [51] rewrote, in an accessible, pedagogical way, the parts of Yang [49] that deal with research lines 1 and 3 above (resp. generalizing Neal and Jacot et al.) (and research line 2 is essentially covered by the combination of [50] and [51]). The main

[6] A coordinatewise nonlinearity is, in the simplest case, a function that is applied to each coordinate of an input vector independently. More generally, see `Nonlin`.

technical result of each of [50] and [51] is a *Master Theorem* that tells one how to mechanically track the correlation between vectors computed in a program. The difference between their versions of the Master Theorems is in the generality of programs considered: In [50], the Master Theorem applies to programs that allow one to re-use matrices, but one cannot use both a matrix and its transpose at the same time. For example, it applies to the computation $AAv$ but not to $A^\top Av$ (both computations reuse the matrix $A$ but the latter do so through its transpose $A^\top$). The Master Theorem of [51] allows $A^\top$, but only under some restrictive condition. For example, it 1) allows $DA^\top u$ where $D$ is a diagonal matrix with diagonal given by $Av$, and $u$ and $v$ are independent standard Gaussian vectors, but 2) disallows $A^\top Av$.

Nevertheless, the restrictions of [50] and [51]'s Master Theorems are natural for their respective settings. Thus, [50] and [51] are able to generalize research lines 1 and 3 to their most practically relevant scenarios (which is more than enough for the purpose of deep learning), but not all practically *conceivable* scenarios (for example, when a weight matrix and its transpose are both used in the forward pass of a neural network). Furthermore, their Master Theorems are *not enough* for proving the random matrix results of this paper, which requires machinery that handles programs like $A^\top Av$. A major contribution of this paper is *to prove a Master Theorem for any Tensor Program*, without the restrictions above[7]. From this would naturally follow the aforementioned generalizations of [50, 51] as well as all of the RMT results in this paper.

**The *Tensor Programs* Series** This paper will complete the rewriting of Yang [49]. Between [50], [51], and this paper, most results of Yang [49] are now re-presented in an accessible way. This paper also finishes the foundation for the current version of the Tensor Programs machinery, which we will rely on crucially for several future papers. We intend this paper to be an authoritative reference for the technical details and the formulation of this foundation, going forward.

**Summary of Our Contributions** 1. We prove the unrestricted Master Theorem, as explained above (Section 2). 2. Give new proofs of the semicircle and Marchenko-Pastur laws, benchmarking our theoretical framework against these fundamental mathematical results (Section 3 and Appendix H). 3. Prove the Free Independence Principle (FIP) (Section 4). 4. Apply FIP to rigorously compute the Jacobian singular value distribution of a randomly initialized NN (Section 5).

For readers familiar with [50, 51], we also: 5. Generalize the GP (research line 1) and NTK (research line 3) results of [50, 51] by allowing the transpose of any weight matrix in the NN forward computation (Appendices B and C). 6. Use FIP to give a new proof of how *Simple GIA Check* allows one to assume GIA when computing NTK rigorously (Appendix D).

Taken together with [50, 51], this paper shows the versatility and power of the *Tensor Programs technique*: To calculate (or show the existence of) some limit, one can just express the quantity of concern in a Tensor Program and apply the Master Theorem mechanically.

## 2 NETSOR$\top$: Language for Neural Computation

There are many versions of languages for Tensor Programs, but we shall focus on one version called NETSOR$\top$ here. This entire section is summarized in Fig. 1, which we encourage the reader to regularly consult over the course of this section. Originally, NETSOR$\top$[8] was motivated by a desire to understand the behavior of large (wide) neural networks. For example, a simple $L$-hidden-layer perceptron can be described by alternating applications of nonlinearities and matrix multiplication: For weight matrices $W^1 \in \mathbb{R}^{n\times d}$ and $W^2, \ldots, W^L \in \mathbb{R}^{n\times n}$, and nonlinearity $\phi : \mathbb{R} \to \mathbb{R}$, such a neural network on input $\xi \in \mathbb{R}^d$ is given by $h^1(\xi) = W^1\xi \in \mathbb{R}^n$, and[9]

$$x^l(\xi) = \phi(h^l(\xi)) \in \mathbb{R}^n, \quad h^{l+1}(\xi) = W^{l+1}x^l(\xi) \in \mathbb{R}^n, \quad \text{for } l = 1, \ldots, L-1, \tag{2}$$

[7] A version of this general Master Theorem was already presented in Yang [49]. Here we focus on giving a more organized, pedagogical proof of it as well as a succinct outline.

[8] pronounced *netsert* or *netser-T*. The *ts* is pronounced like in *tsar*. The *or* is pronounced as in tens*or*. Another way of thinking is NETSOR is just *tensor* with *ten* reversed.

[9] Following neural network convention, $\phi(x) = (\phi(x_1), \ldots, \phi(x_n))$.

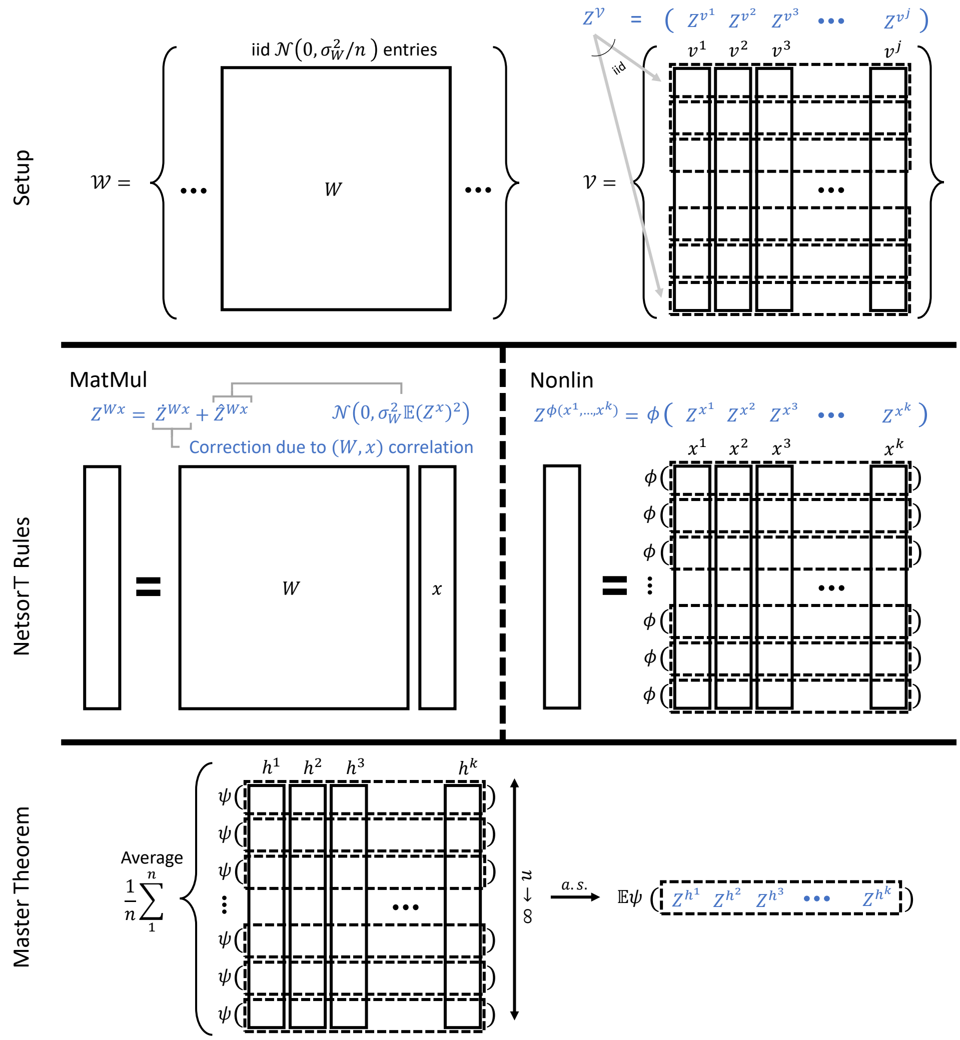


Figure 1: Graphical summary of NETSOR⊤ and its Master Theorem. Vectors $v^i$ are initial vectors of the program, but $x^i$ and $h^i$ can be any vector in the program.

and the network output is $f(\xi) = v^\top x^L(\xi)$ for some weight vector $v \in \mathbb{R}^n$.[10] Thus, intuitively, a language of solely *nonlinearity application* and *matrix multiplication* seems to strike a good balance between 1) simplicity (and ease of analysis), and 2) generality. Indeed, as shown in Yang [50, 51], such a language can express *practically all* of modern deep learning, beyond the toy example here. Yang [51] formalized a version of this language, called NETSOR⊤, which we recall here.

**Definition 2.1.** A NETSOR⊤ program is a sequence of $\mathbb{R}^n$ vectors (which we will refer to as *vectors in the program*) inductively generated via one of the following ways from an initial set $\mathcal{V}$ of random $\mathbb{R}^n$ vectors and a set $\mathcal{W}$ of random $n \times n$ matrices

`Nonlin` Given $\phi : \mathbb{R}^k \to \mathbb{R}$ and $x^1, \ldots, x^k \in \mathbb{R}^n$, we can generate[11] $\phi(x^1, \ldots, x^k) \in \mathbb{R}^n$, where

$$\phi(x^1, \ldots, x^k)_\alpha \stackrel{\text{def}}{=} \phi(x^1_\alpha, \ldots, x^k_\alpha), \quad \text{for each } \alpha \in [n].$$

`MatMul` Given $W \in \mathbb{R}^{n\times n}$ and $x \in \mathbb{R}^n$, we can generate $Wx \in \mathbb{R}^n$ or $W^\top x \in \mathbb{R}^n$.

[10]For simplicity, we omit biases and set equal the widths of all layers, but these two simplifications can easily be removed; see Appendix F.

[11]Again, $\phi$ is applied coordinatewise. Here $\phi$ and $k$ should be thought of as fixed while $n \to \infty$.

For example, in Eq. (2), $h^l(\xi) = W^l x^{l-1}(\xi)$ is an instance of `MatMul`, while $x^l(\xi) = \phi(h^l(\xi))$ is an instance of `Nonlin`. We are interested in understanding the behavior of NETSOR⊤ programs when $n$ is large, and when $\mathcal{V}$ and $\mathcal{W}$ are sampled as follows:

**Setup 2.2** (NETSOR⊤)**.** *1) For each initial $W \in \mathcal{W}$, we sample iid $W_{\alpha\beta} \sim \mathcal{N}(0, \sigma_W^2/n)$ for some variance $\sigma_W^2$ associated to $W$, independent of other $W' \in \mathcal{W}$; 2) for some multivariate Gaussian $Z^{\mathcal{V}} = \left\{Z^h : h \in \mathcal{V}\right\} \in \mathbb{R}^{\mathcal{V}}$, we sample the initial set of vectors $\mathcal{V}$ like $\{h_\alpha : h \in \mathcal{V}\} \sim Z^{\mathcal{V}}$ iid for each $\alpha \in [n]$.*

*Example* 2.3. For example, we will often be interested in understanding the behavior of Eq. (2) when $W^1, \ldots, W^L$ are sampled randomly. In this case, $\mathcal{W}$ consists of $W^2, \ldots, W^L \in \mathbb{R}^{n\times n}$, each of which is sampled like $W^l_{\alpha\beta} \sim \mathcal{N}(0, 1/n)$. On the other hand, $\mathcal{V}$ consists of $h^1(\xi) = W^1\xi$, which is distributed like $h^1(\xi)_\alpha \overset{\mathrm{d}}{=} \mathcal{N}(0, \|\xi\|^2/d)$ if $W^1 \in \mathbb{R}^{n\times d}$ is sampled like $W^1_{\alpha\beta} \sim \mathcal{N}(0, 1/d)$.[12] If we compute the MLP in Eq. (2) on other inputs $\xi_1, \ldots, \xi_k$ as well, then $\mathcal{V} = \{W^1\xi, W^1\xi_1, \ldots, W^1\xi_k\}$ with $Z^{\mathcal{V}}$ being the multivariate Gaussian with covariance $\mathrm{Cov}(Z^{W^1\xi_i}, Z^{W^1\xi_j}) = \xi_i^\top \xi_j/m$.

As discussed in Section 1.2, to understand a random matrix, it suffices to understand random vectors calculated from it. How should we understand the random vectors in a NETSOR⊤ program as $n \to \infty$? As hinted in Yang [50, 51], it turns out that every vector will have roughly iid coordinates in this limit, even though matrix multiplication by $W$ or $W^\top$ will introduce correlation between coordinates for any finite $n$. Following Yang [51], we shall define in Box 1 a random variable $Z^h$ that describes this asymptotic coordinate distribution of each vector $h$; but first, let's use a few examples to motivate the rules governing $Z^h$.

*Example* 2.4. In Example 2.3, by Setup 2.2, we already have $Z^{h^1(\xi)} = \mathcal{N}(0, \|\xi\|^2/n)$, reflecting the fact that each coordinate $h^1(\xi)_\alpha$ is an iid sample of $\mathcal{N}(0, \|\xi\|^2/n)$. For brevity, we drop the explicit dependence of $h$ and $x$ on $\xi$ below. Next, since $x^1 = \phi(h^1)$, we intuitively have $Z^{x^1} = \phi(Z^{h^1})$, as $\phi$ is applied to each coordinate separately.

Now, $h^2 = W^2x^1$ has approximately iid Gaussian coordinates, due to $W^2$ being sampled independent of $x^1$ and a central limit argument. By some simple back-of-the-envelope calculation, the Gaussian coordinates should asymptotically have zero mean and variance $\mathbb{E}(Z^{x^1})^2$, and they are uncorrelated with $x^1$. Thus, it makes sense to set $Z^{h^2} = \mathcal{N}(0, \mathbb{E}(Z^{x^1})^2)$, independent from $Z^{x^1}$ and $Z^{h^1}$.

Our reasoning above can be repeated, and we derive, recursively, that $Z^{h^l} = \mathcal{N}(0, \mathbb{E}(Z^{x^{l-1}})^2)$, $Z^{x^l} = \phi(Z^{h^l})$, with $Z^{h^l}, Z^{x^l}$ independent from $Z^{h^r}, Z^{x^r}$ for all $r < l$.

In this example, the derivation of the $Z$s seems like a fairly simple repackaging of the central limit theorem. However, when both a matrix $A$ and its transpose get involved, the derivation can become complex and subtle quickly, as we show below.

*Example* 2.5. This is already apparent in the $A^\top Av$ example in the introduction, where $v \in \mathbb{R}^n$ is a standard Gaussian vector, and $A \in \mathbb{R}^{n\times n}$, $A_{\alpha\beta} \sim \mathcal{N}(0, 1/n)$ is iid and independent from $v$. By Eq. (1), $A^\top Av$ is roughly $v + g$ for some standard Gaussian vector $g$ independent from $v$. Therefore, we should set $Z^{A^\top Av} = Z^v + G$ for a standard Gaussian $G \in \mathbb{R}$ independent from $Z^v$.

Compare this with $\tilde{A}Av$ for some independent copy $\tilde{A}$ of $A^\top$. Then the calculation of Example 2.4 would have $Z^{\tilde{A}Av}$ be a standard Gaussian independent from $Z^v$. This shows the derivation of the $Z$s can require nuance, depending on the interaction between a matrix and its transpose.

It will turn out that, for any vector $u$ that may depend on $A$ and $A^\top$, $A^\top u$ is always a sum of an asymptotically Gaussian part and a correction term. The Gaussian part comes from assuming $A$ to be independent from $A^\top$ and applying a central limit argument as in Example 2.4. The correction term captures the interaction between $A$ and $A^\top$. For instance, if we take $u = Av$ in the $A^\top Av$ example above, then the Gaussian part is $g$ and the correction term is $v$. In contrast, $\tilde{A}Av$ doesn't have the correction term because $\tilde{A}$ and $A$ are independent.

[12] In neural network settings, we are interested in the limit where the intermediate dimensions $n \to \infty$ but the input dimension $d$ stays fixed. Thus, we treat the input-to-hidden matrix $W^1$ differently from other $W^l$.

In the formal definition (Box 1) of $Z$s, this decomposition of $A^\top u$ is reflected in the definition of $Z^{A^\top u}$ as a sum $\dot{Z}^{A^\top u} + \hat{Z}^{A^\top u}$. Here $\hat{Z}^{A^\top u}$ represents the Gaussian part and $\dot{Z}^{A^\top u}$ represents the correction.

Finally, we state the formal definition of $Z$. Here, the `ZNonlin` and `ZHat` rules are probably intuitive given the examples above, but the `ZDot` rule may appear cryptic at first. We digest the `ZDot` rule more slowly in Remark 2.9 after Box 1.

**Box 1 Key Takeaways for Understanding a NETSOR$\top$ Program**

Each vector $h$ will have coordinates roughly distributed as some random variable $Z^h \in \mathbb{R}$ (in a sense to be formalized in Theorem 2.10), which are *symbolically defined* recursively as:

`ZInit` If $h \in \mathcal{V}$, then $Z^h$ is defined as the random variable given in Setup 2.2. We also set $\hat{Z}^h \stackrel{\text{def}}{=} Z^h$ and $\dot{Z}^h \stackrel{\text{def}}{=} 0$.

`ZNonlin` For any fixed (i.e. constant as $n \to \infty$) $k$ and $\phi : \mathbb{R}^k \to \mathbb{R}$, we have

$$Z^{\phi(x^1,\ldots,x^k)} \stackrel{\text{def}}{=} \phi(Z^{x^1}, \ldots, Z^{x^k}).$$

`ZMatMul` $Z^{Wx} \stackrel{\text{def}}{=} \hat{Z}^{Wx} + \dot{Z}^{Wx}$ for every $W_{\alpha\beta} \sim \mathcal{N}(0, \sigma_W^2/n)$ and vector $x$, where

`ZHat` $\hat{Z}^{Wx}$ is a Gaussian variable with zero mean. Let $\mathcal{V}_W$ denote the set of all vectors in the program of the form $Wy$ for some $y$. Then $\{\hat{Z}^{Wy} : Wy \in \mathcal{V}_W\}$ is defined to be jointly Gaussian with zero mean and covariance

$$\mathrm{Cov}\left(\hat{Z}^{Wx}, \hat{Z}^{Wy}\right) \stackrel{\text{def}}{=} \sigma_W^2 \mathbb{E}\, Z^x Z^y, \quad \text{for any } Wx, Wy \in \mathcal{V}_W.$$

Furthermore, $\{\hat{Z}^{Wy} : Wy \in \mathcal{V}_W\}$ is mutually independent from $\{\hat{Z}^v : v \in \mathcal{V} \cup \bigcup_{\bar{W} \neq W} \mathcal{V}_{\bar{W}}\}$, where $\bar{W}$ ranges over $\mathcal{W} \cup \{A^\top : A \in \mathcal{W}\}$.

`ZDot` By the definition in this box, $Z^x$ is always a deterministic function of a set of $\hat{Z}^\bullet$ random variables. Then the partial derivative $\partial Z^x / \partial \hat{Z}^\bullet$ can be defined symbolically and is another random variable. Then we set

$$\dot{Z}^{Wx} \stackrel{\text{def}}{=} \sigma_W^2 \sum Z^y \,\mathbb{E}\, \frac{\partial Z^x}{\partial \hat{Z}^{W^\top y}},$$

summing over $\hat{Z}^{W^\top y}, W^\top y \in \mathcal{V}_{W^\top}$, that $Z^x$ is a function of (where $\mathcal{V}_{W^\top}$ is defined in `ZHat`). There is some nuiance in this definition, so see Remark 2.11 and 2.12.

The rules above are largely the same as in [51], except the definition of $\hat{Z}^{Wx}$ and $\dot{Z}^{Wx}$. In the restricted NETSOR$\top$ programs of [51], $\dot{Z}^{Wx}$ turns out to be 0 (see Theorem D.1), so $\hat{Z}^{Wx}$ was implicitly identified with $Z^{Wx}$. However, a general NETSOR$\top$ program will have $\dot{Z}^{Wx} \neq 0$.

*Remark* 2.6. Note that $Z^x, \hat{Z}^x, \dot{Z}^x$ only depend on how $x$ is computed in the NETSOR$\top$ program (i.e. the program structure), not the specific (random) value of $x$.

*Example* 2.7. In Example 2.4, $\dot{Z}^{h^l}$ as defined in `ZDot` is always zero (because we never use $W^{l\top}$ in Eq. (2)), and by `ZMatMul`, we simply have $Z^{h^1} = \hat{Z}^{h^1} = \mathcal{N}(0, \|\xi\|^2/d)$, and recursively, $Z^{h^l} = \hat{Z}^{h^l} = \mathcal{N}(0, \mathbb{E}(Z^{x^{l-1}})^2), Z^{x^l} = \phi(Z^{h^l})$, exactly as derived in Example 2.4.

*Example* 2.8. Write $x = A^\top Av$ in Example 2.5 as an explicit NETSOR$\top$ program $y = Av, x = A^\top y$. Then by `ZHat`, $\hat{Z}^x = \mathcal{N}(0, 1)$, independent from $Z^y$ and $Z^v$. By `ZDot`, $\dot{Z}^x = Z^v \,\mathbb{E}\, \frac{\partial Z^y}{\partial \hat{Z}^{Av}} = Z^v \,\mathbb{E}\, \frac{\partial(\hat{Z}^{Av} + \dot{Z}^{Av})}{\partial \hat{Z}^{Av}} = Z^v \,\mathbb{E}\, 1 = Z^v$. This verifies the decomposition $A^\top Av \approx v + g$ in Example 2.5.

*Remark* 2.9 (Intuition for definition of $Z^h$). The rules `ZInit` and `ZNonlin` aptly follow the stated intuition of "$Z^h$ as coordinate distribution of $h$." The `ZMatMul` rule is more complex, so let's digest it a bit.

First suppose $W^\top$ is never used in the program. Then $\dot{Z}^{Wx}$ vanishes, and $Z^{Wx} = \hat{Z}^{Wx}$. The definition of $\hat{Z}^{Wx}$ then roughly says $Wx$ is an isotropic Gaussian vector, which is correlated with other vectors of the form $W\bar{x}$ in a natural way: $\langle Wx, W\bar{x}\rangle \approx \sigma_W^2\langle x, \bar{x}\rangle$, where $\langle,\rangle$ denotes dot product.

However, if $W^\top$ is used previously, then this "Gaussian description" of $Wx$ needs a correction. The definition of $\dot{Z}^{Wx}$ says this correction is a linear combination of previous vectors $y^i$ such that $W^\top y^i$ has been used to compute $x$.

Let us illustrate the meaning of the coefficients of this linear combination through some simple calculations. If $W \in \mathbb{R}^{n\times n}$, $W_{\alpha\beta} \sim \mathcal{N}(0, 1/n)$, and $x \in \mathbb{R}^n$, then $Wx$ should be correlated with $W$. If $x$ depends on $W^\top y^i$ for a collection of vectors $\{y^i\}_i$, then we can detect this correlation as follows: with $\langle,\rangle$ denoting dot product, for each $i$,

$$\langle y^i, Wx\rangle = \langle W^\top y^i, x\rangle.$$

Thus, if $x$ is correlated with $W^\top y^i$, then $Wx$ should also be correlated with $y^i$. The `ZDot` rule, remarkably, says that such correlations are the *only* correction needed to the "Gaussian description" of $Wx$: $Wx$ splits into the sum of 1) a component with coordinates $\approx \dot{Z}^{Wx}$ that resides in the linear span of $\{y^i\}_i$, which is entirely determined by the inner products $\langle y^i, Wx\rangle = \langle W^\top y^i, x\rangle$ for all $i$,[13] and 2) another orthogonal component with coordinates $\approx \hat{Z}^{Wx}$ that comes from naively assuming $W^\top$ to be independent of $W$.

The following *Master Theorem* rigorously relates the *symbolically constructed* random variables $Z^h$ to the vectors $h$ in the program and their *analytic limits*. It is one of our main results and will also be our main workhorse in this paper.

**Theorem 2.10** (NETSOR⊤ Master Theorem)**.** *Fix a* NETSOR⊤ *program. Suppose the initial matrices $\mathcal{W}$ and vectors $\mathcal{V}$ are sampled in the fashion of Setup 2.2. Assume all nonlinearities $\phi$ used in* `Nonlin` *are polynomially bounded. Then for any fixed $k$ and any* polynomially bounded[14] $\psi : \mathbb{R}^k \to \mathbb{R}$*, as $n \to \infty$,*

$$\frac{1}{n}\sum_{\alpha=1}^{n}\psi(h_\alpha^1, \ldots, h_\alpha^k) \xrightarrow{\text{a.s.}} \mathbb{E}\,\psi(Z^{h^1}, \ldots, Z^{h^k}), \tag{3}$$

*for any collection of vectors $h^1, \ldots, h^k$ in the program, where $Z^{h^i}$ are defined in Box 1.*[15]

Theorem 2.10 says that each "coordinate slice" $(h_\alpha^1, \ldots, h_\alpha^k)$ can be thought of as an iid copy of $(Z^{h^1}, \ldots, Z^{h^k})$. Indeed, as consequences of the Master Theorem, Theorems A.5 and A.6 formalize this intuition into a convergence in distribution. Versions of Theorem 2.10 with convergence in mean (instead of almost sure) are also available (Theorems A.1 and A.2). Theorem 2.10 strictly generalizes the Master Theorem in Yang [51]. The NETSOR Master Theorem in Yang [50] allows nonlinearities growing faster than polynomially, but otherwise is also a special case of Theorem 2.10. We outline the proof of Theorem 2.10 in Section 6 and give a full proof in Appendix L.

*Remark* 2.11 (Partial derivative). The partial derivative in `ZDot` should be interpreted as follows. By a simple inductive argument, $Z^x$ for every vector $x$ in the program is defined *uniquely* as a deterministic function $\varphi(\hat{Z}^{x^1}, \ldots, \hat{Z}^{x^k})$ of some $x^1, \ldots, x^k$ in $\mathcal{V}$ or introduced by `MatMul`.[16] For instance, in Example 2.8, $Z^x = \hat{Z}^x + \hat{Z}^v$ if $v \in \mathcal{V}$, so $\varphi$ is given by $\varphi(a, b) = a + b$. Then

$$\partial Z^x/\partial \hat{Z}^{x^i} \stackrel{\text{def}}{=} \partial_i\varphi(\hat{Z}^{x^1}, \ldots, \hat{Z}^{x^k}), \quad \text{and} \quad \partial Z^x/\partial \hat{Z}^z \stackrel{\text{def}}{=} 0 \text{ for any } z \notin \{x^1, \ldots, x^k\}.$$

Note this definition depends on the precise way the program is written, not just on the underlying mathematics. For example, if $y, z \in \mathcal{V}$ and $x = \phi(W(y+z))$, then $Z^x = \phi(\hat{Z}^{W(y+z)})$ so that $\partial Z^x/\partial \hat{Z}^{Wy} = \partial Z^x/\partial \hat{Z}^{Wz} = 0$. If instead, we have $x = \phi(Wy + Wz)$, then $Z^x = \phi(\hat{Z}^{Wy} + \hat{Z}^{Wz})$ so that $\partial Z^x/\partial \hat{Z}^{W(x+y)} = 0$. However, in both cases, $\dot{Z}^{W^\top x} = (Z^y + Z^z)\,\mathbb{E}\,\phi'(\hat{Z}^{W(y+z)})$.

[13] The set of coefficients $\{\frac{1}{n}\langle W^\top y^i, x\rangle \approx \mathbb{E}\, Z^{W^\top y^i} Z^x\}_i$ is linearly related to the coefficients $\{\mathbb{E}\,\hat{Z}^{W^\top y^i} Z^x\}_i$, as can be seen from an easy inductive argument. The latter is then linearly related to the partial derivative expectations of `ZDot` by Remark 2.12.

[14] $\phi : \mathbb{R}^k \to \mathbb{R}$ is *polynomially-bounded* if $|\phi(x)| \le C\|x\|^p + c$ for some $p, C, c > 0$, for all $x \in \mathbb{R}^k$.

[15] Difference with [49, Thm 6.3]: We have gotten rid of the "rank convergence" assumption (which we call "rank stability" in this paper) by showing that it comes for free. See Theorem 6.3 and see CoreSet and Lemma L.11 in Appendix L.

[16] In the context of NETSOR⊤$^+$ introduced later in Appendix E, this is still true, but $\varphi$ here will take the form of a parametrized nonlinearity $\varphi(-; \mathring{\theta}_1, \ldots, \mathring{\theta}_l)$ with some deterministic parameters $\mathring{\theta}_1, \ldots, \mathring{\theta}_l$.

*Remark* 2.12 (Partial derivative expectation). The quantity $\mathbb{E}\,\frac{\partial Z^x}{\partial \hat{Z}^{W^\top y}}$ is well defined if $Z^x$ is differentiable in $\hat{Z}^{W^\top y}$. However, even if this is not the case, e.g. if $x = \theta(W^\top y)$ where $\theta$ is the Heavyside step function, we can still define this expectation by leveraging Stein's lemma (Lemma K.4):

In ZDot, suppose $\{W^\top y^i\}_{i=1}^k$ are all elements of $\mathcal{V}_{W^\top}$ introduced before $x$. Define the matrix $C \in \mathbb{R}^{k\times k}$ by $C_{ij} \stackrel{\text{def}}{=} \mathbb{E}\, Z^{y^i} Z^{y^j}$ and define the vector $b \in \mathbb{R}^k$ by $b_i \stackrel{\text{def}}{=} \mathbb{E}\, \hat{Z}^{W^\top y^i} Z^x$. If $a = C^+ b$ (where $C^+$ denotes the pseudoinverse of $C$), then in ZDot we may set

$$\sigma_W^2\,\mathbb{E}\,\frac{\partial Z^x}{\partial \hat{Z}^{W^\top y^i}} = a_i. \tag{4}$$

This definition agrees with the partial derivative expectation by Stein's lemma (Lemma K.4) when the latter is well defined. Theorem 2.10 holds with this broader definition of partial derivative expectation.

**Extensions** In Appendix E, we describe NETSOR$\top^+$, an extension to NETSOR$\top$ by allowing programs to compute the average coordinate of a vector, and use such scalars in Nonlin. In Appendix F, we also describe modification to the Master Theorems if, instead of requiring all matrices in $\mathcal{W}$ to be square, we allow rectangular matrices.

## 3 Semicircle Law

The *Semicircle Law* [45, 46] is a classical result, of central importance in statistics, physics, and engineering [8, 20, 29, 44], on the spectrum of a random Hermitian matrix with independent, zero-mean entries. It says that, as the size of the matrix tends to infinity, the distribution of its eigenvalues tends to a *semicircle distribution.*

**Definition 3.1.** The semicircle distribution $\mu_{\mathrm{sc}}$ is the distribution with density $\propto \sqrt{4-x^2}$.

**Theorem 3.2** (Semicircle Law for GOE)**.** *For each $n \geq 1$, define the random symmetric matrix*[17] $A = A(n) = W + W^\top$ *for iid Gaussian matrix $W \in \mathbb{R}^{n\times n}$, $W_{\alpha\beta} \sim \mathcal{N}(0, 1/2n)$. Let $\lambda_1, \ldots, \lambda_n \in \mathbb{R}$ be the eigenvalues of $A$; these are random variables. Then for every compactly supported, continuous $\varphi : \mathbb{R} \to \mathbb{R}$, as $n \to \infty$, we have*

$$\frac{1}{n}\sum_{\alpha=1}^{n} \varphi(\lambda_\alpha) \xrightarrow{\text{a.s.}} \mathop{\mathbb{E}}_{\lambda\sim\mu_{\mathrm{sc}}} \varphi(\lambda).$$

In this section, we give a new proof of this beautiful result using the NETSOR$\top$ Master Theorem. Our purpose is two-fold: we 1) give concrete examples of how to compute with the Master Theorem, especially the new $\dot{Z}^{Wx}$ rule, and 2) demonstrate our framework is at least powerful enough to prove this cornerstone result. In Appendix H, we also prove the *Marchenko-Pastur Law* with this technique.

By the well-known *Moment Method* (see Fact G.2), it suffices to prove

$$n^{-1}\operatorname{tr} A^r \xrightarrow{\text{a.s.}} \mathop{\mathbb{E}}_{\lambda\sim\mu_{\mathrm{sc}}} \lambda^r, \quad r = 1, 2, \ldots$$

It is well known that the semicircle distribution has moments given by the Catalan numbers $C_k$

$$\mathop{\mathbb{E}}_{\lambda\sim\mu_{\mathrm{sc}}} \lambda^r = \begin{cases} C_k & \text{if } r = 2k \\ 0 & \text{otherwise} \end{cases}$$

where the Catalan numbers $C_k$ are the unique numbers satisfying

$$C_0 = 1, \quad C_{k+1} = \sum_{i=0}^{k} C_i C_{k-i}. \tag{5}$$

The first few Catalan numbers are $C_0 = 1, C_1 = 1, C_2 = 2, C_3 = 5, C_4 = 14$. For more background, see Tao [42].

[17] This is known as the Gaussian Orthogonal Ensemble (GOE). While we will only talk about the GOE case here, the semicircle law holds for generic random matrices with iid entries. We hope to show universality of NETSOR$\top$ Master Theorem in the future, which would automatically generalize our proof here to such cases.

## 3.1 The Main Proof Idea

We need to show for all integers $k$,

$$n^{-1}\operatorname{tr} A^{2k} \xrightarrow{\text{a.s.}} C_k, \quad \text{and} \quad n^{-1}\operatorname{tr} A^{2k+1} \xrightarrow{\text{a.s.}} 0.$$

To do so, we use a trivial but useful equality: for any $M \in \mathbb{R}^{n\times n}$

$$\operatorname{tr} M = \mathop{\mathbb{E}}_{z} z^\top M z, \quad \text{for} \quad z \sim \mathcal{N}(0, I). \tag{6}$$

Then

$$\operatorname{tr} A^{2k} = \mathop{\mathbb{E}}_{z} z^\top A^{2k} z,$$

which we can express as a NETSOR$\top$ program: Let $\mathcal{V} = \{z\}$, and $\mathcal{W} = \{W\}$ (with sampling data $Z^z = \mathcal{N}(0, 1), \sigma_W^2 = 1/2$). With $z^0 = z$, we define recursively

$$\boxed{x^t = W z^{t-1}, \quad y^t = W^\top z^{t-1}, \quad z^t = x^t + y^t.} \tag{7}$$

Then, mathematically, we have computed

$$z^t = A^t z, \quad \text{and thus} \quad \operatorname{tr} A^t = \mathop{\mathbb{E}}_{z} z^\top z^t.$$

Note $A^t$ is the $t$th power of $A$ but the $t$ in $x^t, y^t, z^t$ appear as indices. By the Master Theorem (and some additional arguments below), it then suffices to show

$$\mathbb{E}\, Z^z Z^{z^{2k}} = C_k,\ \mathbb{E}\, Z^z Z^{z^{2k+1}} = 0 \quad \text{so that} \quad n^{-1}\mathop{\mathbb{E}}_{z} z^\top z^{2k} \xrightarrow{\text{a.s.}} C_k,\ n^{-1}\mathop{\mathbb{E}}_{z} z^\top z^{2k+1} \xrightarrow{\text{a.s.}} 0$$

## 3.2 Examples of First Few Moments

We first construct the random variables $Z^{z^t}, Z^{x^t}, Z^{y^t}$, for $z^t, x^t, y^t$ defined in Eq. (7). While we can do the proof much more succinctly, in the pedagogical spirit, let's do the first few manually to get a feel for it.

**First Moment** First we have $Z^{z^0} = \mathcal{N}(0, 1)$ by definition. Then $Z^{x^1} = \hat{Z}^{x^1} + \dot{Z}^{x^1} = \hat{Z}^{x^1}$ because we have not used $W^\top$ yet, so $\dot{Z}^{x^1} = 0$. Now, by ZHat, $\mathbb{E}(\hat{Z}^{x^1})^2 = \sigma_W^2\, \mathbb{E}(Z^{z^0})^2 = 1/2 \cdot 1 = 1/2$. Therefore, $Z^{x^1} = \hat{Z}^{x^1}$ is a Gaussian with zero mean and variance $1/2$, independent from $Z^{z^0}$.

A similar reasoning shows $\hat{Z}^{y^1} = \mathcal{N}(0, 1/2)$ as well, independent from $Z^{z^0}$ and $Z^{x^1}$. Next we apply ZDot to calculate $\dot{Z}^{y^1}$. But $y^1$ does not depend on any vector of the form $W^\top \bullet$, so $\dot{Z}^{y^1} = 0$. Thus, $Z^{y^1} = \hat{Z}^{y^1}$, as a summary,

$$Z^{z^0} = \mathcal{N}(0, 1), \quad Z^{x^1} = \mathcal{N}(0, 1/2), \quad Z^{y^1} = \mathcal{N}(0, 1/2),$$

all independent from each other. In general, $Z^{y^i}$ and $Z^{x^i}$ are symmetric, as illustrated here. Then

$$Z^{z^1} = Z^{x^1} + Z^{y^1} \implies \mathbb{E}\left(Z^{z^1}\right)^2 = 1 \quad \text{and} \quad n^{-1}\operatorname{tr} A \xrightarrow{\text{a.s.}} \mathbb{E}\, Z^{z^1} Z^{z^0} = 0.$$

**Second Moment** Next,

$$Z^{x^2} = \hat{Z}^{x^2} + \dot{Z}^{x^2}$$

with $\hat{Z}^{x^2}$ independent from $Z^{y^1}, Z^z$, but jointly Gaussian with $\hat{Z}^{x^1} = Z^{x^1}$ with (co-)variance

$$\mathbb{E}(\hat{Z}^{x^2})^2 = \frac{1}{2}\mathbb{E}(Z^{z^1})^2 = \frac{1}{2}, \quad \operatorname{Cov}(\hat{Z}^{x^2}, \hat{Z}^{x^1}) = \frac{1}{2}\mathbb{E}\, Z^{z^1} Z^{z^0} = 0 \quad \text{(i.e. independent)}.$$

On the other hand, since $y^1 = W^\top z^0$ is the only usage of $W^\top$ in defining $x^2$,

$$\dot{Z}^{x^2} = \frac{1}{2} Z^{z^0}\, \mathbb{E}\frac{\partial Z^{z^1}}{\partial \hat{Z}^{y^1}} = \frac{1}{2} Z^{z^0}\, \mathbb{E}\, 1 = \frac{1}{2} Z^{z^0}.$$

Altogether, we have

$$Z^{x^2} = \hat{Z}^{x^2} + \dot{Z}^{x^2} = \hat{Z}^{x^2} + \frac{1}{2} Z^{z^0}, \quad \text{and symmetrically,} \quad Z^{y^2} = \hat{Z}^{y^2} + \dot{Z}^{y^2} = \hat{Z}^{y^2} + \frac{1}{2} Z^{z^0}.$$

Since $\hat{Z}^{x^2}$ and $\hat{Z}^{y^2}$ are zero-mean and independent from $Z^{z^0}$, we have

$$Z^{z^2} = Z^{x^2} + Z^{y^2} = \hat{Z}^{x^2} + \hat{Z}^{y^2} + Z^{z^0} \implies n^{-1}\operatorname{tr} A^2 \xrightarrow{\text{a.s.}} \mathbb{E}\, Z^{z^2} Z^{z^0} = \mathbb{E}(Z^{z^0})^2 = 1 = C_1.$$

### 3.3 Proof for General Moments

**Overview** Notice how $Z^{z^t}, t = 1, 2$, above are of the form $Z^{Z^t} = \tau_t Z^{z^0} + S$ for some $\tau_t \in \mathbb{R}$ and some zero-mean $S$ independent from $Z^{z^0}$. We will show this is the case for all $Z^{z^t}$, so that $\mathbb{E}\, Z^{z^t} Z^{z^0} = \mathbb{E}(\tau_t Z^{z^0} + S) Z^{z^0} = \tau_t$. By Theorem 3.3 below (which is just the Master Theorem plus a standard truncation argument to turn the almost sure convergence into almost sure convergence of conditional expectations), this means $n^{-1} \operatorname{tr} A^t = n^{-1} \mathbb{E}_z\, z^\top z^t \xrightarrow{\text{a.s.}} \tau_t$. Finally, we will show $\tau_{2k+1} = 0$ and $\tau_{2k}$ satisfies the same recurrence as $C_k$, which then yields the desired result.

The following theorem is a direct consequence of the more general Theorem A.2.

**Theorem 3.3.** *With the same premise as in Theorem 2.10, suppose further $\psi$ is quadratically bounded and all nonlinearities used in* `Nonlin` *are linearly bounded. Let $\mathcal{S} \subseteq \mathcal{V}$ be a subcollection of initial vectors. Then*

$$\frac{1}{n} \mathop{\mathbb{E}}_{\mathcal{S}} \sum_{\alpha=1}^{n} \psi(h^1_\alpha, \ldots, h^k_\alpha) \xrightarrow{\text{a.s.}} \mathbb{E}\, \psi(Z^{h^1}, \ldots, Z^{h^k}). \tag{8}$$

*where $\mathbb{E}_{\mathcal{S}}$ denotes conditional expectation given the values of vectors in $\mathcal{V} \setminus \mathcal{S}$ and of matrices in $\mathcal{W}$.*

**Calculations** In general, $\{\hat{Z}^{x^s}\}_s$, $\{\hat{Z}^{y^s}\}_s$, $Z^{z^0}$ are zero-mean and independent from one another. We have $\{\hat{Z}^{x^s}\}_s$ is jointly Gaussian with covariance

$$\operatorname{Cov}(\hat{Z}^{x^s}, \hat{Z}^{x^r}) = \frac{1}{2}\, \mathbb{E}\, Z^{z^{s-1}} Z^{z^{r-1}}$$

and $\{\hat{Z}^{y^s}\}_s$ satisfies symmetric covariance identities. In addition, by `ZDot`,

$$\dot{Z}^{x^{t+1}} = \frac{1}{2} \sum_{s=0}^{t-1} Z^{z^s}\, \mathbb{E}\, \frac{\partial Z^{z^t}}{\partial \hat{Z}^{y^{s+1}}}, \quad \dot{Z}^{y^{t+1}} = \frac{1}{2} \sum_{s=0}^{t-1} Z^{z^s}\, \mathbb{E}\, \frac{\partial Z^{z^t}}{\partial \hat{Z}^{x^{s+1}}}$$

It's easy to see that there are deterministic coefficients $b^t_s \in \mathbb{R}$ (independent of $n$) so that

$$Z^{z^t} = \sum_{s=1}^{t} b^t_s (\hat{Z}^{x^s} + \hat{Z}^{y^s}) + b^t_0 Z^{z^0}. \tag{9}$$

Then

$$n^{-1} \operatorname{tr} A^t \xrightarrow{\text{a.s.}} \mathbb{E}\, Z^{z^t} Z^{z^0} = b^t_0.$$

Note

$$\frac{\partial Z^{z^t}}{\partial \hat{Z}^{x^s}} = \frac{\partial Z^{z^t}}{\partial \hat{Z}^{y^s}} = b^t_s.$$

Consequently,

$$\dot{Z}^{x^{t+1}} = \dot{Z}^{y^{t+1}} = \frac{1}{2} \sum_{s=0}^{t-1} Z^{z^s} b^t_{s+1}$$

and

$$\begin{aligned}
Z^{z^{t+1}} &= \hat{Z}^{x^{t+1}} + \hat{Z}^{y^{t+1}} + \dot{Z}^{x^{t+1}} + \dot{Z}^{y^{t+1}} = \hat{Z}^{x^{t+1}} + \hat{Z}^{y^{t+1}} + \sum_{s=0}^{t-1} Z^{z^s} b^t_{s+1} \\
&= \hat{Z}^{x^{t+1}} + \hat{Z}^{y^{t+1}} + \sum_{s=0}^{t-1} b^t_{s+1} \left( \sum_{r=1}^{s} b^s_r (\hat{Z}^{x^r} + \hat{Z}^{y^r}) + b^s_0 Z^{z^0} \right) \\
&= \hat{Z}^{x^{t+1}} + \hat{Z}^{y^{t+1}} + \sum_{r=1}^{t-1} (\hat{Z}^{x^r} + \hat{Z}^{y^r}) \sum_{s=r}^{t-1} b^s_r b^t_{s+1} + \sum_{s=0}^{t-1} b^s_0 b^t_{s+1} Z^{z^0}
\end{aligned}$$

Matching coefficients with Eq. (9), this implies

$$b^{t+1}_{t+1} = 1, \quad b^{t+1}_r = \sum_{s=r}^{t-1} b^s_r b^t_{s+1}, \forall r \leq t-1.$$

Using the Catalan identity Eq. (5), we can check that the solution is

$$b_r^t = \begin{cases} C_{(t-r)/2} & \text{if } t-r \text{ is even} \\ 0 & \text{otherwise.} \end{cases}$$

Then

$$n^{-1} \operatorname{tr} A^{2k} \xrightarrow{\text{a.s.}} b_0^{2k} = C_k, \quad n^{-1} \operatorname{tr} A^{2k+1} \xrightarrow{\text{a.s.}} b_0^{2k+1} = 0$$

as desired.

### 3.4 Comparison with the Classical Proof

The classical way of calculating the expected moments $\mathbb{E} \operatorname{tr} A^k$ is to expand this trace as a sum of products $A_{\alpha_1\alpha_2} \cdots A_{\alpha_k\alpha_1}$ of entries of $A$, and then notice that all summands contribute vanishingly other than those that have each unique term $A_{\alpha_i\alpha_{i+1}}$ appear with power 2. Then the computation of $\mathbb{E} \operatorname{tr} A^k$ to top order can be seen to boil down to a counting problem of *non-crossing partitions*. Since the solution of such counting problem is exactly the Catalan numbers, we have the desired result.

In contrast to this classical proof, our proof by Tensor Programs is purely symbolic (i.e. does not require manually identifying leading and subleading terms and doing the combinatorics). The role of the mathematician here has been mostly to express the moment computation as a NETSOR$\top$ program, and the rest follows from the Master Theorem (and can be done by a computer). While perhaps after unwinding, this technique may be implicitly doing a sort of combinatorics similar to the classical proof, we believe this particular way of *repackaging* is useful. Indeed, when the matrix ensemble in question involves nonlinear dependencies, such as in a neural network Jacobian, our techique applies readily (see Section 5), while the classical proof is hard to transfer as we can no longer expand the matrix power in terms of the matrix entries because of the nonlinear dependencies[18].

**Universality** On the other hand, a current drawback of this Tensor Programs proof is the limitation of the Master Theorem to Gaussian matrices. However, we expect universality will hold in our case, and the Master Theorem can be proven for general, iid matrices, for example through some version of the Lindeberg Replacement Trick. We leave this to future work.

## 4 The Free Independence Principle of NETSOR$\top$ Programs

A powerful analogue of *independence* (of scalar random variables) in random matrix theory is called *Asymptotic Free Independence*, or *Asymptotic Freeness*, defined below in Definition 4.1.

**Definition 4.1.** Fix $k$. Consider collections of random matrices $\mathcal{W}_n^1, \ldots, \mathcal{W}_n^k \subseteq \mathbb{R}^{n\times n}$ for each $n \geq 1$, of constant cardinalities (with $n$). (For example, each $\mathcal{W}_n^i$ can be $\{W, W^\top\}$ for some weight matrix $W$ in a neural network). We say $\mathcal{W}_n^1, \ldots, \mathcal{W}_n^k$ are *almost surely asymptotically free*[19], if

$$n^{-1} \operatorname{tr}\left(\prod_{i=1}^{k} \left(P_i(\mathcal{W}_n^{j_i}) - \tau_i\right)\right) \xrightarrow{\text{a.s.}} 0$$

where $\tau_i = n^{-1} \operatorname{tr}(P_i(\mathcal{W}_n^{j_i}))$, $P_i$ is a (noncommutative) polynomial in $|\mathcal{W}_n^{j_i}|$ variables, and $j_1, \ldots, j_k \in [k]$ are indices with no two adjacent $j_i$ equal, with $\{P_i\}_i$, $\{j_i\}_i$ independent of $n$.[20] [21]

[18]One can still naively expand through the nonlinearities by Taylor expansion, but 1) this requires the nonlinearities to be smooth, and 2) this asks for significant effort for the mathematician to count and bound the lower order terms. In contrast, the proof by Tensor Programs will be, for the most part, mechanical, following the Master Theorem, and allows nonsmooth nonlinearities.

[19]We can also speak of asymptotically free in expectation, in which case, we want the expectation of the trace to converge to 0. In most scenarios (where we have tail bounds on the matrix operator norms), this is a weaker notion than almost sure asymptotic freeness.

[20]Note here $\prod$ is a non-commutative product, where in its expansion, $i$ goes from right to left.

[21]One can also formulate the asymptotic freeness of subalgebras of a non-commutative algebra, in which case our definition here is equivalent to the freeness of the subalgebras generated by the respective collections of random matrices.

Whereas the independence of *scalar* random variables allows one to compute the expectation of their sum and product easily, the asymptotic freeness of random *matrices* allows one to compute the asymptotic spectral distributions of their sum and (matrix) product easily. Independence implies that two random variables are in "general position" and thus cannot conspire to fluctuate in the same direction. Likewise, asymptotic freeness of two random matrices, intuitively, implies that their respective eigenvectors and eigenvalues lie in general position to each other, so that their spectral distributions would combine in predictable ways were they to be summed or multiplied. For more background, see [42].

*A priori*, because the activations and preactivations of a neural network depend in a highly nonlinear and complex manner in the weight matrices (and biases), one can hardly suspect that activations can be "independent" from the weight matrices in some way. However, we will show, in a randomly initialized neural network of any architecture, the weight matrices are asymptotically free from (the diagonal matrices formed from) the preactivations of the network. This will follow from the much more general *Free Independence Principle* of NETSOR$\top$ programs, Theorem 4.2. We give a proof in Appendix J that follows the same overall strategy as the proof of the Semicircle Law in Section 3.

**Theorem 4.2** (Free Independence Principle, for Tensor Programs)**.** *Consider any* NETSOR$\top$ *program $\pi$ with polynomially bounded nonlinearities and Setup 2.2. Then the random matrix collections $\{W, W^\top\}$ for every $W \in \mathcal{W}$, along with the collection of diagonal matrices $\mathcal{D}(\pi)$ (defined immediately below in Definition 4.3) are asymptotically free as $n \to \infty$.*

**Definition 4.3.** Suppose $\mathcal{X}$ is a subset of ($\mathbb{R}^n$) vectors in a program. Let $\mathcal{D}(\mathcal{X})$ denote the (infinite) collection of diagonal matrices formed from bounded, coordinatewise images of $\mathcal{X}$:

$$\mathcal{D}(\mathcal{X}) \stackrel{\text{def}}{=} \{\mathrm{Diag}(\psi(x^1, \ldots, x^k)) : k \geq 0;\;\; x^1, \ldots, x^k \in \mathcal{X};\;\; \psi : \mathbb{R}^k \to \mathbb{R} \text{ bounded}\} \subseteq \mathbb{R}^{n \times n}. \tag{10}$$

If $\pi$ is a program, we write $\mathcal{D}(\pi)$ to denote $\mathcal{D}(\{\text{all vectors in } \pi\})$.

Note $\psi$ in Eq. (10) is distinct from the nonlinearities in the program $\pi$. For example, if $\pi$ expresses the forward pass of a ReLU MLP, then $\pi$ has unbounded nonlinearity (ReLU) but $\psi$ in Eq. (10) can be the step function, which is bounded. We have kept $\psi$ bounded in Eq. (10) for technical reasons (similar to the appearance of *bounded* continuous functions in the definition of convergence in distribution). But we believe Theorem 4.2 holds when $\psi$ is more generally polynomially bounded. This generalization would follow if Theorem 2.10 holds for almost sure convergence of conditional means; see Conjecture A.4.

**Intuition and Discussion** One can perhaps accept NETSOR$\top$ programs as a formalization of "reasonable ways" to compute vectors (and their diagonal matrices) from a set of random matrices. Then FIP says that a random Gaussian matrix is asymptotically free from any diagonal matrix that "depends on it in a reasonable way." This formalizes the intuition that singular vectors of Gaussian matrices are in general position to singular vectors of diagonal matrices, so one may expect these matrices to be asymptotically free.

We shall see next that Theorem 4.2 allows us to easily compute the asymptotic Jacobian singular value distribution of a randomly initialized neural network.

**Extension to NETSOR$\top^+$ programs with variable dimensions** Theorem 4.2 holds as stated for NETSOR$\top$ programs with variable dimensions. It also holds for NETSOR$\top^+$ programs with variable dimensions if nonlinearities are parameter-controlled and rank stability (Assumption E.7) is satisfied.

## 5 Jacobian Singular Values of a Randomly Initialized Neural Network

**Notation** We denote the empirical spectral distribution of a random matrix $W$ by $\mu_W$.[22] We write $\mu_W \boxtimes \mu_V$ to denote the free multiplicative convolution of $\mu_W$ and $\mu_V$.[23]

**Review of semirigorous computation of Jacobian singular value distribution in prior works.** Analyses in previous works [36] are mostly semirigorous and proceed, for example, as follows:

[22] i.e. $\mu_W = \frac{1}{n}\sum_{\alpha=1}^{n} \delta_{\lambda_\alpha}$, where $\delta_{\lambda_\alpha}$ is the Dirac Delta distribution centered on the $\alpha$th eigenvalue $\lambda_\alpha$.

[23] If $W$ and $V$ are asymptotically free random matrices, then $\mu_W \boxtimes \mu_V$ converges to the asymptotic spectral distribution of $WV$. See Speicher [41] for more details on free probability.

If $f(\xi)$ is an MLP as in Eq. (2) with width $n$, then the Jacobian $J = \partial h^L/\partial h^1 \in \mathbb{R}^{n\times n}$, on a fixed input $\xi$, can be written as[24]

$$J = W^L D^{L-1} W^{L-1} D^{L-2} \cdots W^2 D^1,$$

where $W^l$ are its weight matrices and $D^l = \mathrm{Diag}(\phi'(h^l))$ are the diagonal matrices with activation derivatives on the diagonals. Then the singular values of $J$ are the square roots of the spectrum of

$$J^\top J = D^1 W^{2\top} \cdots D^{L-1} W^{L\top} W^L D^{L-1} \cdots W^2 D^1.$$

Now here's the non-rigorous part: With the random initialization $W^1_{\alpha\beta} \sim \mathcal{N}(0, 1/d)$ and $W^2_{\alpha\beta}, \ldots, W^L_{\alpha\beta} \sim \mathcal{N}(0, 1/n)$, prior works assume that the random matrix collections

$$\{W^2, W^{2\top}\}, \ldots, \{W^L, W^{L\top}\}, \{D^1\}, \ldots, \{D^{L-1}\} \text{ are asymptotically free.} \tag{11}$$

Then, with this assumption, noting that the spectrum of $AB$ and $BA$ agree for any two matrices $A, B$ of appropriate sizes[25], we have

$$\mu_{J^\top J} = \mu_{D^1 W^{2\top} \ldots D^{L-1} W^{L\top} W^L D^{L-1} \ldots W^2 D^1} = \mu_{(D^1)^2 W^{2\top} \ldots D^{L-1} W^{L\top} W^L D^{L-1} \ldots W^2}$$

so that, by the freeness asumption of $D^1$ from the other matrices, we have

$$\lim_{n\to\infty} \mu_{J^\top J} = \lim_{n\to\infty} \mu_{(D^1)^2} \boxtimes \lim_{n\to\infty} \mu_{W^2 \ldots D^{L-1} W^{L\top} W^L D^{L-1} \ldots W^2}$$

where $\boxtimes$ denotes multiplicative free convolution. Repeating this logic yields

$$\lim_{n\to\infty} \mu_{J^\top J} = \lim_{n\to\infty} \mu_{(D^1)^2} \boxtimes \lim_{n\to\infty} \mu_{W^2 W^{2\top}} \boxtimes \cdots \boxtimes \lim_{n\to\infty} \mu_{W^L W^{L\top}}.$$

Since $\lim_{n\to\infty} \mu_{W^i W^{i\top}}$ is just the Marchenko-Pastur distribution (Eq. (26)) and $\lim_{n\to\infty} \mu_{(D^l)^2}$ is (at least, at the time, heuristically) distributed like $\phi'(Z^{h^l})$, the standard S-transform technique (see Speicher [41]) allows one to explicitly compute $\lim_{n\to\infty} \mu_{J^\top J}$.

**Our contribution** Of course, by Theorem 4.2 and the NETSOR$\top$ program Eq. (2), Eq. (11) is now completely rigorous. In fact, Theorem 4.2 implies a much more general result.

**Corollary 5.1** (Free Independence Principle)**.** *In any randomly initialized neural network expressible in* NETSOR$\top$ *program $\pi$ with polynomially bounded* `Nonlin`*, the weight matrices are asymptotically free from (the diagonal matrices formed from) bounded images of all preactivations: the random weight matrix collections $\{W, W^\top\}$ for matrices $W \in \mathcal{W}$, along with $\mathcal{D}(\pi)$ (defined in Definition 4.3), are asymptotically free. Furthermore, for any mutually independent partition $\mathcal{X}_1, \ldots, \mathcal{X}_k$ of $\{Z^x : x \in \pi\}$, the random diagonal matrix collections $\mathcal{D}(\{x : Z^x \in \mathcal{X}_1\}), \ldots, \mathcal{D}(\{x : Z^x \in \mathcal{X}_k\})$ are asymptotically free.*

Since almost all neural networks can be written in NETSOR$\top$, as shown in Yang [50, 51], Corollary 5.1 is a very universal result. By this corollary, the entire computation above in fact (after easily checking $Z^{h^1}, \ldots, Z^{h^{L-1}}$ are mutually independent) yields a proof of the following almost sure convergence:

**Theorem 5.2.** *Consider an MLP $f(\xi)$ as in Eq. (2) with $L$ hidden layers, width $n$, and nonlinearity $\phi$ with bounded weak derivative[26] $\phi'$. Then its Jacobian $J = \partial h^L/\partial h^1 \in \mathbb{R}^{n\times n}$ on a fixed input $\xi \in \mathbb{R}^d$ has the $n \to \infty$ limit*

$$\mu_{J^\top J} \xrightarrow{\text{a.s.}} \mu_{\phi'(Z^{h^1})} \boxtimes \cdots \boxtimes \mu_{\phi'(Z^{h^{L-1}})} \boxtimes \mu_{\text{mp}}^{\boxtimes(L-1)}$$

*where $\mu_{\text{mp}}$ is the Marchenko-Pastur distribution with shape ratio 1 (see Eq. (26)), $\mu_{\phi'(Z^{h^l})}$ is the distribution of the random variable $\phi'(Z^{h^l})$, and $\xrightarrow{\text{a.s.}}$ denotes almost sure convergence of random measures (Definition G.1).*

Most nonlinearities $\phi$ in deep learning practice have bounded weak derivatives (such as tanh or ReLU), and thus is covered by Theorem 5.2. When $\phi$ is identity, Theorem 5.2 just recovers the Marchenko-Pastur Law. Note here the dependence of $\mu_{J^\top J}$ on input $\xi$ comes purely through $\mu_{\phi'(Z^{h^1})}, \ldots, \mu_{\phi'(Z^{h^{L-1}})}$.

[24]Note we study the Jacobian of the *body* of the network, whose dimensions $n$ tend to infinity. The perhaps more natural input-output Jacobian has finite dimensions, so there's no asymptotic distribution to speak of.

[25]This can be seen by applying Sylvester's Determinant Theorem $\det(zI - AB) = \det(zI - BA)$ on the characteristic polynomials of $AB$ and $BA$.

[26]i.e. $\phi$ is almost everywhere differentiable and $\phi'$ is a function that agrees with this derivative almost everywhere. ReLU is a typical example of $\phi$, whose weak derivative is the step function and is bounded.

**Generalizations** Theorem 5.2 generalizes straightforwardly to the case when the MLP has non-unit width ratios, in which case the $\mu_{\text{mp}}^{\boxtimes(L-1)}$ should be replaced by free multiplicative convolution of Marchenko-Pastur distributions of different shape ratios (see Eq. (26)). Also, like the remark below Theorem 4.2, we expect Theorem 5.2 holds if $\phi'$ is just polynomially bounded and Corollary 5.1 holds if $\mathcal{D}$ is defined using polynomially bounded (instead of bounded) $\psi$. We leave this to future work.

**Computing the Jacobian Singular Value Distribution of Any Neural Architecture** More generally, Corollary 5.1 allows us to compute the Jacobian singular value of neural networks of any architecture (such as residual networks, recurrent networks, convolutional networks, etc).

Indeed, the Jacobian can always be expressed as a polynomial in the matrices and (the diagonal matrices formed from) vectors of an appropriate NETSOR$\top$ program. By Corollary 5.1, these random matrix collections are asymptotically free. The asymptotic spectral distribution of such a polynomial in asymptotically free matrices can be computed via operator-valued free probability [30, Chapter 10]. Thus our result here yields a general algorithm for computing the asymptotic singular value distribution of the Jacobian of a randomly initialized neural network of any architecture. Alternatively, one can always resort to explicit moment computations via Tensor Programs, as in Section 3.

**Simple GIA Check Implies GIA** It turns out that the Neural Tangent Kernel depends on the computation of backpropagation only through quantities that can be expressed as 2nd moment of Jacobian singular values, if Simple GIA Check (Condition 1) holds. By the results of this section, we thus can replace the matrices $W^\top$ in backpropagation with copies that are independent from $W$ the forward pass. See Appendix D for more details.

# 6 Proof Sketch of the Master Theorem

Here we explain the main ideas of the proof of Theorem 2.10. The Master Theorems of Yang [50, 51] all use some subset of these ideas for their proofs, so this section also summarizes the core insights there. We start by proving an easier version of Theorem 2.10 that assumes all nonlinearities in `Nonlin` are sufficiently smooth (Section 6.1). Then we show how to remove the smoothness assumption (Sections 6.2 and 6.2.2). Finally, we give an outline in Section 6.3 for proving the form of $\dot{Z}$ in Box 1.

## 6.1 Master Theorem Proof Sketch with Sufficient Regularity Assumptions

A natural intuition for proving the Master Theorem is to perform induction on the number of vectors. Suppose further for simplicity that all matrices $W \in \mathcal{W}, W \in \mathbb{R}^{n\times n}$ are sampled like $W_{\alpha\beta} \sim \mathcal{N}(0, 1/n)$ . Here we sketch the proof of the simpler statement that Eq. (3) converges to *some* limit, when all nonlinearities $\phi$ used in `Nonlin` are smooth enough. We will discuss the form of the limit itself in another section (Section 6.3). Under these assumptions, the core idea is similar to the proof of Bayati and Montanari [6] for Approximate Mesage Passing (but we will not require knowledge of this proof below).

**Definition 6.1.** We say a vector is a *G-var* if it is in $\mathcal{V}$ or it is introduced by `MatMul`.[27]

Suppose the program has G-vars $g^1, \ldots, g^m$ introduced in that order. Since all other vectors can be expressed as a `Nonlin` image of them, it suffices to prove Eq. (3) for $g^1, \ldots, g^m$, i.e. for any *sufficiently smooth* $\psi : \mathbb{R}^m \to \mathbb{R}$, we have

$$\frac{1}{n}\sum_{\alpha=1}^{n} \psi(g^1_\alpha, \ldots, g^m_\alpha) \xrightarrow{\text{a.s.}} \mathbb{E}\,\psi(Z^{g^1}, \ldots, Z^{g^m}). \tag{12}$$

The base case is when $g^1, \ldots, g^m \in \mathcal{V}$ are the initial vectors. Then Eq. (12) converges by law of large numbers.

[27]The letter *G* elicits the intuition that the vector is roughly *Gaussian* plus a correction term. We will not use the terms *A-var* and *H-var* from Yang [49, 50, 51], in favor of more intuitive terms *matrix* and *vector* in the program.

For the inductive case, suppose $g^{m+1} = Ah$, where $h = \phi(g^1, \dots, g^m)$, for some *sufficiently smooth* $\phi : \mathbb{R}^m \to \mathbb{R}$, and WLOG for $A \in \mathcal{W}$. We want to show Eq. (12) for $m \leftarrow m+1$. The *key idea is to condition* $A$ on $g^1, \dots, g^m$, and then try to reduce Eq. (12) for $m+1$ to Eq. (12) for $m$ through a law of large numbers on the remaining randomness in $g^{m+1}$. This conditioning puts a linear constraint on $A$ in the form of

$$\mathbf{X} = A\mathbf{Y}, \quad \mathbf{U} = A^\top \mathbf{V}$$

for matrices $\mathbf{X}, \mathbf{Y}, \mathbf{U}, \mathbf{V}$ with previous vectors as columns. For example, if the program is $\{g^2 = Ag^1, g^3 = A^\top g^2\}$, then conditioning on $g^1, g^2, g^3$, we have $\mathbf{X} = g^2, \mathbf{Y} = g^1, \mathbf{U} = g^3, \mathbf{V} = g^2$. By standard formulas for Gaussian conditioning (see Lemma K.9), we can derive the following conditional distribution[28]

$$A \overset{\mathrm{d}}{=}_{g^1,\dots,g^m} E + \Pi_1 \tilde{A} \Pi_2, \quad g^{m+1} \overset{\mathrm{d}}{=}_{g^1,\dots,g^m} \left(E + \Pi_1 \tilde{A} \Pi_2\right) h,$$

where $\tilde{A}$ is an iid copy of $A$, $E \in \mathbb{R}^{n \times n}$ is the "conditional mean", and $\Pi_1, \Pi_2 \in \mathbb{R}^{n \times n}$ are two orthogonal projection matrices into subspaces of dimension $n - O(1)$. Then we can see each coordinate $g^{m+1}_\alpha$ is conditionally distributed like $(Eh)_\alpha + \sigma \Pi_1 \zeta$ where $\sigma^2 = \|\Pi_2 h\|^2/n$ and $\zeta \sim \mathcal{N}(0, I)$. Now we make several approximations that can be made rigorous using the smoothness of $\psi$:

1. Since $\Pi_1$ has small corank, we approximate $\Pi_1 \approx I$.[29]
2. It turns out $\sigma^2$ can be rewritten as a continuous function of quantities in the form of Eq. (12), so by induction hypothesis, $\sigma \xrightarrow{\text{a.s.}} \mathring{\sigma}$ for some deterministic limit $\mathring{\sigma} \geq 0$ (see Lemma L.6). We make the approximation $\sigma \approx \mathring{\sigma}$.
3. Similarly, it turns out $Eh = \sum_{i=1}^r a_i h^i$ where $h^i$ are some previous vectors (which are necessarily `Nonlin` images of $g^1, \dots, g^m$), and each $a_i$ is a (continuous function of) average of Eq. (12)'s form (see Lemma L.7). They converge $a_i \xrightarrow{\text{a.s.}} \mathring{a}_i$ by induction hypothesis, so we approximate $a_i \approx \mathring{a}_i$. Noet that the specific forms of $a_i$ and $h^i$ here will dictate the form of $\dot{Z}^{g^{m+1}}$; see Section 6.3.

In summary, we have now approximated, for each $\alpha \in [n]$,

$$g^{m+1}_\alpha \overset{d}{\approx}_{g^1,\dots,g^m} \sum_{i=1}^r \mathring{a}_i h^i_\alpha + \mathring{\sigma} \zeta_\alpha, \quad \zeta_\alpha \sim \mathcal{N}(0,1). \tag{13}$$

Thus, as $g^{m+1}_\alpha$ is iid in $\alpha \in [n]$ with this approximation, by law of large numbers, we should expect Eq. (12) for $m \leftarrow m+1$ to concentrate around its conditional expectation for large $n$:

$$\left| \frac{1}{n} \sum_{\alpha=1}^n \psi(g^1_\alpha, \dots, g^m_\alpha, g^{m+1}_\alpha) - S \right| \xrightarrow{\text{a.s.}} 0,$$
$$\text{where} \quad S \overset{\text{def}}{=} \frac{1}{n} \sum_{\alpha=1}^n \mathop{\mathbb{E}}_{z \sim \mathcal{N}(0,1)} \psi \left( g^1_\alpha, \dots, g^m_\alpha, \mathring{\sigma} z + \sum_{i=1}^r \mathring{a}_i h^i_\alpha \right).$$

Now $S$ is in the form of Eq. (12) so we can apply induction hypothesis. This finishes the proof sketch.

In this proof sketch, we have substituted many quantities for their limits (that are inductively proven to exist). This is only possibly because we assume all nonlinearities are sufficiently smooth. What if we don't have this assumption? (This is important, for example, for expressing the backpropagation of a ReLU neural network, since ReLU's (weak) derivative is not continuous).

[28] see Appendix K.1 for definition of $\overset{\mathrm{d}}{=}$

[29] this is roughly because $\Pi_1$ is multiplied to $\tilde{A}$, which generically sends a vanishing amount of its image to the subspace represented by $\Pi_1$.

### 6.2 Getting Rid of Smoothness Assumption

The key insight into the removal of smoothness assumption on `Nonlin` is that, *we get smoothness for free from averaging*. Here's the main strategy: If we can show the conditional concentration

$$\left|\frac{1}{n}\sum_{\alpha=1}^{n}\psi(g_\alpha^1,\ldots,g_\alpha^m,g_\alpha^{m+1})-\bar{S}\right| \xrightarrow{\text{a.s.}} 0,$$

$$\text{where}\quad \bar{S}\stackrel{\text{def}}{=}\mathbb{E}\left[\frac{1}{n}\sum_{\alpha=1}^{n}\psi(g_\alpha^1,\ldots,g_\alpha^m,g_\alpha^{m+1})\middle| g^1,\ldots,g^m\right], \tag{14}$$

then $\bar{S}$ can be expressed as an average of *smooth* functions of $g_\alpha^1,\ldots,g_\alpha^m$ and some scalars $a_1,\ldots a_r,\sigma$ (as in Items 2 and 3 above) that converge to deterministic limits. The smoothness of these functions come from the Gaussian averaging inside the conditional expectation, even if $\psi$ itself is not smooth. This works as long as $\sigma>0$; we shall discuss this assumption more thoroughly in Section 6.2.2. With this smoothness, we can again replace $a_1,\ldots,a_r,\sigma$ with their limits $\mathring{a}_1,\ldots,\mathring{a}_r,\mathring{\sigma}$, and then the proof is finished by the induction hypothesis, as in Section 6.1.

Thus, if we can prove Eq. (14) without using smoothness of $\psi$, then we would be done. Revisiting the approximations we made in our first-attempt proof, we see that we need to remake our law of large number argument (Item 1) without substituting $\Pi_1\approx I$. This substitution has been quite convenient, as it made $g_\alpha^{m+1}$ conditionally iid in $\alpha$, so that the usual law of large numbers can apply. Now, we have to manually wrestle with the correlations between $g_\alpha^{m+1}$ and $g_\beta^{m+1}$ for pairs $\alpha,\beta\in[n]$.

#### 6.2.1 Law of Large Numbers for Images of Weakly Correlated Gaussians

We thus prove a new law of large numbers for this case. It says that the average of images of weakly correlated Gaussians will converge deterministically.

**Theorem 6.2** (LLN for Images of Weakly Correlated Gaussians (Simplified))**.** *Consider a triangular array $\{\zeta_1^n,\ldots,\zeta_n^n\}_{n\ge1}$ of Gaussian variables, where each row is given by $\zeta^n\sim\mathcal{N}(0,\Sigma^n)$ and the covariance matrix $\Sigma^n$ satisfies $\sum_{\alpha\neq\beta}(\Sigma_{\alpha\beta}^n)^2/(\Sigma_{\alpha\alpha}^n\Sigma_{\beta\beta}^n)=O(1)$. Consider any polynomially-bounded $\phi:\mathbb{R}\to\mathbb{R}$. Then the triangular array $\{\phi(\zeta_1^n),\ldots,\phi(\zeta_n^n)\}_{n\ge1}$ satisfies a strong law of large numbers.*[30]

This theorem will be applicable to $\Sigma^n=\Pi_1$ as $\Pi_1$ is a projection matrix with low corank and can therefore be seen to have small off-diagonal entries (see Remark K.20). This would then finish the proof of Eq. (14) without assuming smoothness of `Nonlin`.

Let us then sketch a proof of Theorem 6.2. It is instructive to first show a *weak* law of large numbers (Corollary K.22) by bounding the variance of the fluctuation around the mean (Theorem K.21). Suppose, for simplicity, $\phi$ is even, so that $\mathbb{E}\,\phi(\zeta_\alpha^n)=0$. Then we need to bound $\mathbb{E}\left(\frac{1}{n}\sum_{\alpha=1}^{n}\phi(\zeta_\alpha^n)\right)^2$. Expanding the square, we have $n$ diagonal terms $\frac{1}{n^2}\mathbb{E}\,\phi(\zeta_\alpha^n)^2,\alpha\in[n]$, and $n(n-1)$ cross terms $\frac{1}{n^2}\mathbb{E}\,\phi(\zeta_\alpha^n)\phi(\zeta_\beta^n),\alpha\neq\beta$. The former contributes $O(1/n)$. We shall show the latter is, too.

Let $\phi(x)=b_1H_1(x)+b_2H_2(x)+\cdots$ be the Hermite expansion of $\phi$, with $H_i$ denoting the $i$th Hermite polynomial. Note that there's no $b_0$ term because $\phi$ has mean 0. Then a neat identity Fact K.13 says, for any jointly Gaussian $(z_1,z_2)$ with zero-mean, unit variance, and covariance $c$, we have

$$\mathbb{E}\,\phi(z_1)\phi(z_2)=b_1^2c+b_2^2c^2+\cdots. \tag{15}$$

If we assume $\Sigma^n$ has unit diagonal, then by this identity, we have

$$\frac{1}{n^2}\sum_{\alpha\neq\beta}\mathbb{E}\,\phi(\zeta_\alpha^n)\phi(\zeta_\beta^n)=\frac{1}{n^2}\sum_{i\ge1}b_i^2\sum_{\alpha\neq\beta}\left(\Sigma_{\alpha\beta}^n\right)^i\stackrel{1}{\le}\frac{1}{n^2}\sum_{i\ge1}b_i^2 n\sqrt{\sum_{\alpha\neq\beta}\left(\Sigma_{\alpha\beta}^n\right)^{2i}}$$

$$\stackrel{2}{\le}\frac{1}{n}\left(\sum_{i\ge1}b_i^2\right)\sqrt{\sum_{\alpha\neq\beta}\left(\Sigma_{\alpha\beta}^n\right)^2}\stackrel{3,4}{=}O(1/n).$$

[30]The full theorem (Corollary K.24) allows each $\zeta_\alpha^n$ to have its own $\phi_\alpha:\mathbb{R}\to\mathbb{R}$ and this is in fact what's needed to finish the proof of Theorem 2.10. However, for conveying the main insights, we will be content with the simplified statement here.

Here we used 1) power mean inequality, 2) $\Sigma^n_{\alpha\beta} \leq 1 \implies (\Sigma^n_{\alpha\beta})^{2i} \leq (\Sigma^n_{\alpha\beta})^2$ for $i \geq 1$, 3) $\sum_{i\geq 1} b_i^2 = \mathbb{E}_{z\sim\mathcal{N}(0,1)} \phi(z)^2 = O(1)$, and 4) $\sum_{\alpha\neq\beta} \left(\Sigma^n_{\alpha\beta}\right)^2 = O(1)$ by assumption. This finishes the proof sketch of the weak version of Theorem 6.2 under the simplifying assumptions of even $\phi$ and $\Sigma^n$ having unit diagonal (which can be removed easily by complicating the proof a bit).

To prove the strong version of Theorem 6.2, we need to bound the higher moments $\mathbb{E}\left(\frac{1}{n}\sum_{\alpha=1}^n \phi(\zeta^n_\alpha)\right)^p, p \geq 4$. This requires a similar analysis as the above, but much more technically involved. For example, Eq. (15) generalizes to higher-order cross terms $\mathbb{E}\,\phi(z_1)\cdots\phi(z_k)$ for jointly Gaussian $(z_1,\ldots,z_k)$, but the resulting expression Theorem K.15 is difficult to use directy. Instead, we need to divide into cases and bound it: either all pairs $(z_i, z_j)$ have uniformly weak correlations (Lemma K.16), or some pair has really large correlation comparatively (Lemma K.17). For full details, check Theorem K.23.

#### 6.2.2 Rank Stability

We have glossed over two important points in the above sketches: A) The scalars $\sigma$ and $a_i$ in Eq. (13) depend on the pseudo-inverse of some Gram matrix of vectors in the program. Even though this Gram matrix will converge almost surely, its rank could drop suddenly in the limit, causing its pseudo-inverse to diverge, so that $\sigma$ and $a_i$ also diverge (see Proposition L.5). B) if $\sigma = 0$ (the "conditional standard deviation of $g^{m+1}$ given $g^1,\ldots,g^m$"), then the conditional expectation involves no "averaging" so the argument of "getting smoothness for free" does not work (see Appendix L.6). More precisely, there are several scenarios for $\sigma$ and its limit $\mathring{\sigma}$:

1. $\mathring{\sigma} > 0$ so, since $\sigma \xrightarrow{\text{a.s.}} \mathring{\sigma}$, $\sigma > 0$ almost surely as well.
2. $\mathring{\sigma} = 0$ and $\sigma = 0$ a.s. for large $n$.
3. $\mathring{\sigma} = 0$ and $\sigma > 0$ a.s., only converging to 0 at $n = \infty$.

In the case of 1), the arguments of Section 6.2 go through, but this most likely won't work in the cases of 2) and 3). Intuitively, case 2) can perhaps allow us to reduce $g^{m+1}$ to a deterministic function of $g^1,\ldots,g^m$, and show $\frac{1}{n}\sum_{\alpha=1}^n \psi(g^1_\alpha,\ldots,g^m_\alpha,g^{m+1}_\alpha)$ is a.s. equal to $\frac{1}{n}\sum_{\alpha=1}^n \bar{\psi}(g^1_\alpha,\ldots,g^m_\alpha)$ for an appropriate $\bar{\psi}$, so we can apply the induction hypothesis. But this argument cannot apply to case 3), because of the randomness in $g^{m+1}$ even after conditioning.

It turns out the intuition for case 2) is correct and case 3) doesn't happen! This will follow from the property of rank stability, which will simultaneously solve both problem A) and B):

**Theorem 6.3** (Rank Stability)**.** *Let $y, y^1,\ldots,y^k$ be any collection of vectors in a* NETSOR$\top$ *program. If $Z^y = b_1 Z^{y^1} + \ldots + b_k Z^{y^k}$, then almost surely, for large enough $n$, $y = b_1 y^1 + \cdots + b_k y^k$.*

Indeed, if $\mathring{\sigma} = 0$, then it turns out we can show $Z^{g^{m+1}}$ is a linear combination of $Z^{g^1},\ldots,Z^{g^m}$, so rank stability tells us $g^{m+1}$ is the *same* linear combination of $g^1,\ldots,g^m$ (almost surely, for large $n$), i.e. we are in case 2). Thereafter, we can straightforwardly reduce to induction hypothesis. On the other hand, this also shows case 3) can never occur. Now it just suffices to show Theorem 6.3, which can be done essentially by a simultaneous induction with the main induction hypothesis. See Appendix L for more details.

### 6.3 The Calculation of $\dot{Z}$

As remarked in Remark 2.9, the adjunction relation $\langle y, Wx\rangle = \langle W^\top y, x\rangle$ is importantly related to the existence and form of $\dot{Z}^{Wx}$. If we assume the Master Theorem, then this adjunction implies the identity $\mathbb{E}\, Z^y Z^{Wx} = \mathbb{E}\, Z^{W^\top y} Z^x$ for any $x, y$ in the program. In fact, a very similar identity exists as well for $\dot{Z}$:

$$\mathbb{E}\, Z^y \dot{Z}^{Wx} = \mathbb{E}\, \hat{Z}^{W^\top y} Z^x.$$

This is proved in Lemma L.3. With this identity, we can verify (Lemma L.10) that, in the notation of Eq. (13) (using the precise form of $\mathring{a}_i$ omitted here),

$$\sum_{i=1}^r \mathring{a}_i \dot{Z}^{h^i} = \dot{Z}^{g^{m+1}}.$$

Additionally, straightforward computation (Lemma L.19) shows

$$\sum_{i=1}^{r} \mathring{a}_i \hat{Z}^{h^i} + \mathring{\sigma} z \overset{\mathrm{d}}{=}_{g^1,\ldots,g^m} \hat{Z}^{g^{m+1}}, z \sim \mathcal{N}(0,1).$$

Then we can rewrite Eq. (13) heuristically as

$$\begin{aligned} g_\alpha^{m+1} &\overset{d}{\approx}_{g^1,\ldots,g^m} \sum_{i=1}^{r} \mathring{a}_i h_\alpha^i + \mathring{\sigma}\zeta_\alpha, \quad \zeta_\alpha \sim \mathcal{N}(0,1). \\ &\overset{d}{\approx}_{g^1,\ldots,g^m} \sum_{i=1}^{r} \mathring{a}_i Z^{h^i} + \mathring{\sigma}\zeta_\alpha = \left(\sum_{i=1}^{r} \mathring{a}_i \dot{Z}^{h^i}\right) + \left(\sum_{i=1}^{r} \mathring{a}_i \hat{Z}^{h^i} + \mathring{\sigma}\zeta_\alpha\right) = \dot{Z}^{g^{m+1}} + \hat{Z}^{g^{m+1}}, \end{aligned}$$

as desired.

# 7 Related Works

**Jacobian Singular Value Distribution of a Neural Network** Pennington et al. [36, 37] originally studied the multilayer-perceptron (MLP)'s Jacobian singular value distribution, in the limit of large width. Ling et al. [25], Tarnowski et al. [43] generalized this analysis to residual MLPs and Xiao et al. [48] to convolutional networks. These works assumed certain asymptotic freeness as in Eq. (11), which was first proven rigorously by Yang [49] and presented in a more accessible way here. Recently, Pastur [34] gave a new, direct proof of this asymptotic distribution for MLP by induction in its depth and standard random matrix machinery. In comparison, our technique here is more general, both in the type of architectures allowed (any that is expressible in NETSOR$\top$, which by Yang [50, 51] includes practically all architectures) and in the nonlinearities involved (Pastur [34] assumes $\phi, \phi'$ are both bounded but we only require $\phi'$ is bounded).

In this paper, we are concerned with the input-output Jacobian of a neural network on a fix input. Other works have considered the Jacobian with respect to the parameters, where the Jacobian has dimension #parameters $\times$ #data points (e.g. [35]). This is a related but distinct case from the input-output Jacobian.

**$\dot{Z}^{Wx}$ as the *Onsager Correction Term*** This $\dot{Z}^{Wx}$ has appeared before, in a limited setting, in the literature of asymmetric message passing as the *Onsager correction term*. Asymmetric message passing [10] is an algorithm that tries to recover a ground truth signal $v$ from a noisy measurement $Wv$ of it obtained by a matrix $W$. This algorithm repeatedly multiplies the measurement by $W$ and its transpose $W^\top$, interleaving with coordinatewise nonlinearities. In between matrix multiplications, this Onsager correction term is subtracted explicitly. Bayati and Montanari [6] famously proved the recovery properties of this algorithm as the matrix size tends to infinity. This algorithm can be written down in a NETSOR$\top$ program, and (a version of) Bayati and Montanari [6]'s results can be realized as a corollary of the NETSOR$\top$ Master Theorem (see Yang [49]). In contrast, the Master Theorem keeps track of the Onsager correction term throughout arbitrary computation. This allows us to prove powerful theorems like the Semicircle Law and the Free Independence Principle, as shown in this paper.

**The *Tensor Programs* Series** This paper is the third in the *Tensor Programs* series, following Yang [50, 51]. The unrestricted Master Theorem, the new proofs of semicircle and Marchenko-Pastur laws, and the singular value distribution result for multilayer-perceptron originally appeared in Yang [49], but here we present a clear, pedagogical presentation of them, with some technical improvements as well. The Free Independence Principle is new and unique to this paper. Compared to the Master Theorem of Yang [51], our Master Theorem here does not assume the BP-like condition (which implies that $W^\top$ can be assumed independent from $W$, i.e. GIA), but rather works for *any* NETSOR$\top$ program.

# 8 Conclusion

In this work, we proved a Master Theorem for any NETSOR$\top$ program and applied this new theorem to give new proofs of the semicircle and Marchenko-Pastur laws, as well as to propose and prove the

*Free Independence Principle (FIP)*. FIP then allows us to calculate the asymptotic Jacobian singular value distribution of a neural network. These results suggest a new way to approach nonlinear random matrix theory pertaining to deep learning. More generally, in combination with Yang [50, 51], they demonstrate the versatility of the *Tensor Programs* technique.


## Acknowledgements

We thank Boris Hanin, Sam Schoenholz, Zhiyuan Li, Jeffrey Pennington, Etai Littwin, Ilya Razenshteyn, Bobby He, Ryan O'Donnell, Edward Hu, Michael Santacroce, Jason Lee, Judy Shen, Jascha Sohl-Dickstein for feedback on working copies of this manuscript.


## References


[1] Sina Alemohammad, Zichao Wang, Randall Balestriero, and Richard Baraniuk. The recurrent neural tangent kernel, 2020.

[2] Zeyuan Allen-Zhu, Yuanzhi Li, and Yingyu Liang. Learning and Generalization in Overparameterized Neural Networks, Going Beyond Two Layers. *arXiv:1811.04918 [cs, math, stat]*, November 2018. URL `http://arxiv.org/abs/1811.04918`.

[3] Zeyuan Allen-Zhu, Yuanzhi Li, and Zhao Song. A Convergence Theory for Deep Learning via Over-Parameterization. *arXiv:1811.03962 [cs, math, stat]*, November 2018. URL `http://arxiv.org/abs/1811.03962`.

[4] Zeyuan Allen-Zhu, Yuanzhi Li, and Zhao Song. On the Convergence Rate of Training Recurrent Neural Networks. *arXiv:1810.12065 [cs, math, stat]*, October 2018. URL `http://arxiv.org/abs/1810.12065`.

[5] Sanjeev Arora, Simon S. Du, Wei Hu, Zhiyuan Li, Ruslan Salakhutdinov, and Ruosong Wang. On Exact Computation with an Infinitely Wide Neural Net. *arXiv:1904.11955 [cs, stat]*, April 2019. URL `http://arxiv.org/abs/1904.11955`.

[6] Mohsen Bayati and Andrea Montanari. The dynamics of message passing on dense graphs, with applications to compressed sensing. *IEEE Transactions on Information Theory*, 57(2): 764–785, February 2011. ISSN 0018-9448, 1557-9654. doi: 10.1109/TIT.2010.2094817. URL `http://arxiv.org/abs/1001.3448`.

[7] Minmin Chen, Jeffrey Pennington, and Samuel Schoenholz. Dynamical Isometry and a Mean Field Theory of RNNs: Gating Enables Signal Propagation in Recurrent Neural Networks. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 873–882, Stockholmsmässan, Stockholm Sweden, July 2018. PMLR. URL `http://proceedings.mlr.press/v80/chen18i.html`.

[8] Romain Couillet and Merouane Debbah. *Random Matrix Methods for Wireless Communications*. Cambridge University Press, Cambridge, 2011. ISBN 978-0-511-99474-6. doi: 10.1017/CBO9780511994746. URL `http://ebooks.cambridge.org/ref/id/CBO9780511994746`.

[9] Amit Daniely, Roy Frostig, and Yoram Singer. Toward Deeper Understanding of Neural Networks: The Power of Initialization and a Dual View on Expressivity. In D. D. Lee, M. Sugiyama, U. V. Luxburg, I. Guyon, and R. Garnett, editors, *Advances in Neural Information Processing Systems 29*, pages 2253–2261. Curran Associates, Inc., 2016. URL `http://papers.nips.cc/paper/6427-toward-deeper-understanding-of-neural-networks-the-power-of-initialization-and-a-dual-view-on-expressivity.pdf`.

[10] David L. Donoho, Arian Maleki, and Andrea Montanari. Message Passing Algorithms for Compressed Sensing. *Proceedings of the National Academy of Sciences*, 106(45):18914–18919, November 2009. ISSN 0027-8424, 1091-6490. doi: 10.1073/pnas.0909892106. URL `http://arxiv.org/abs/0907.3574`.

[11] Simon S. Du, Xiyu Zhai, Barnabas Poczos, and Aarti Singh. Gradient Descent Provably Optimizes Over-parameterized Neural Networks. *arXiv:1810.02054 [cs, math, stat]*, October 2018. URL http://arxiv.org/abs/1810.02054.

[12] Simon S. Du, Kangcheng Hou, Barnabás Póczos, Ruslan Salakhutdinov, Ruosong Wang, and Keyulu Xu. Graph neural tangent kernel: Fusing graph neural networks with graph kernels, 2019.

[13] Adrià Garriga-Alonso, Laurence Aitchison, and Carl Edward Rasmussen. Deep Convolutional Networks as shallow Gaussian Processes. *arXiv:1808.05587 [cs, stat]*, August 2018. URL http://arxiv.org/abs/1808.05587.

[14] Boris Hanin. Which Neural Net Architectures Give Rise To Exploding and Vanishing Gradients? January 2018. URL https://arxiv.org/abs/1801.03744.

[15] Boris Hanin and David Rolnick. How to Start Training: The Effect of Initialization and Architecture. *arXiv:1803.01719 [cs, stat]*, March 2018. URL http://arxiv.org/abs/1803.01719.

[16] Soufiane Hayou, Arnaud Doucet, and Judith Rousseau. On the Selection of Initialization and Activation Function for Deep Neural Networks. *arXiv:1805.08266 [cs, stat]*, May 2018. URL http://arxiv.org/abs/1805.08266.

[17] Tamir Hazan and Tommi Jaakkola. Steps Toward Deep Kernel Methods from Infinite Neural Networks. *arXiv:1508.05133 [cs]*, August 2015. URL http://arxiv.org/abs/1508.05133.

[18] Jiri Hron, Yasaman Bahri, Jascha Sohl-Dickstein, and Roman Novak. Infinite attention: Nngp and ntk for deep attention networks, 2020.

[19] Arthur Jacot, Franck Gabriel, and Clément Hongler. Neural Tangent Kernel: Convergence and Generalization in Neural Networks. *arXiv:1806.07572 [cs, math, stat]*, June 2018. URL http://arxiv.org/abs/1806.07572.

[20] Steven Kivelson, Dung-Hai Lee, and Shou-Cheng Zhang. Global phase diagram in the quantum Hall effect. *Physical Review B*, 46(4):2223–2238, July 1992. ISSN 0163-1829, 1095-3795. doi: 10.1103/PhysRevB.46.2223. URL https://link.aps.org/doi/10.1103/PhysRevB.46.2223.

[21] A. M. Lane, R. G. Thomas, and E. P. Wigner. Giant Resonance Interpretation of the Nucleon-Nucleus Interaction. *Physical Review*, 98(3):693–701, May 1955. ISSN 0031-899X. doi: 10.1103/PhysRev.98.693. URL https://link.aps.org/doi/10.1103/PhysRev.98.693.

[22] Nicolas Le Roux and Yoshua Bengio. Continuous neural networks. In *Artificial Intelligence and Statistics*, pages 404–411, 2007.

[23] Jaehoon Lee, Yasaman Bahri, Roman Novak, Sam Schoenholz, Jeffrey Pennington, and Jascha Sohl-dickstein. Deep Neural Networks as Gaussian Processes. In *International Conference on Learning Representations*, 2018. URL https://openreview.net/forum?id=B1EA-M-0Z.

[24] Jaehoon Lee, Lechao Xiao, Samuel S. Schoenholz, Yasaman Bahri, Jascha Sohl-Dickstein, and Jeffrey Pennington. Wide Neural Networks of Any Depth Evolve as Linear Models Under Gradient Descent. *arXiv:1902.06720 [cs, stat]*, February 2019. URL http://arxiv.org/abs/1902.06720.

[25] Zenan Ling, Xing He, and Robert C. Qiu. Spectrum concentration in deep residual learning: a free probability approach. *arXiv:1807.11694 [cs, stat]*, February 2019. URL http://arxiv.org/abs/1807.11694. arXiv: 1807.11694.

[26] Etai Littwin, Tomer Galanti, and Lior Wolf. On the optimization dynamics of wide hypernetworks, 2020.

[27] V A Marčenko and L A Pastur. DISTRIBUTION OF EIGENVALUES FOR SOME SETS OF RANDOM MATRICES. *Mathematics of the USSR-Sbornik*, 1(4):457–483, April 1967. ISSN 0025-5734. doi: 10.1070/SM1967v001n04ABEH001994. URL http://stacks.iop.org/0025-5734/1/i=4/a=A01?key=crossref.1d0b803ddab02373cb6b0690a61e734a.

[28] Alexander G. de G. Matthews, Mark Rowland, Jiri Hron, Richard E. Turner, and Zoubin Ghahramani. Gaussian Process Behaviour in Wide Deep Neural Networks. *arXiv:1804.11271 [cs, stat]*, April 2018. URL http://arxiv.org/abs/1804.11271.

[29] F. Mezzadri and N. C. Snaith, editors. *Recent Perspectives in Random Matrix Theory and Number Theory*. Cambridge University Press, 1 edition, June 2005. ISBN 978-0-521-62058-1 978-0-511-55049-2. doi: 10.1017/CBO9780511550492. URL https://www.cambridge.org/core/product/identifier/9780511550492/type/book.

[30] J.A. Mingo and R. Speicher. *Free Probability and Random Matrices*. Fields Institute Monographs. Springer New York, 2017. ISBN 978-1-4939-6942-5. URL https://books.google.com/books?id=d7wpDwAAQBAJ.

[31] Radford M Neal. *BAYESIAN LEARNING FOR NEURAL NETWORKS*. PhD Thesis, University of Toronto, 1995.

[32] Roman Novak, Lechao Xiao, Jaehoon Lee, Yasaman Bahri, Daniel A Abolafia, Jeffrey Pennington, and Jascha Sohl-Dickstein. Bayesian Deep Convolutional Networks with Many Channels are Gaussian Processes. *arXiv preprint arXiv:1810.05148*, 2018.

[33] Ryan O'Donnell. *Analysis of boolean functions*. Cambridge University Press, New York, NY, 2014. ISBN 978-1-107-03832-5.

[34] Leonid Pastur. On random matrices arising in deep neural networks. gaussian case, 2020.

[35] Jeffrey Pennington and Pratik Worah. The Spectrum of the Fisher Information Matrix of a Single-Hidden-Layer Neural Network. In *Advances in Neural Information Processing Systems 31*, page 10, 2018.

[36] Jeffrey Pennington, Samuel Schoenholz, and Surya Ganguli. Resurrecting the sigmoid in deep learning through dynamical isometry: theory and practice. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 4788–4798. Curran Associates, Inc., 2017. URL http://papers.nips.cc/paper/7064-resurrecting-the-sigmoid-in-deep-learning-through-dynamical-isometry-theory-and-practice.pdf.

[37] Jeffrey Pennington, Samuel S. Schoenholz, and Surya Ganguli. The Emergence of Spectral Universality in Deep Networks. *arXiv:1802.09979 [cs, stat]*, February 2018. URL http://arxiv.org/abs/1802.09979. arXiv: 1802.09979.

[38] George Philipp and Jaime G. Carbonell. The Nonlinearity Coefficient - Predicting Overfitting in Deep Neural Networks. *arXiv:1806.00179 [cs, stat]*, May 2018. URL http://arxiv.org/abs/1806.00179.

[39] Ben Poole, Subhaneil Lahiri, Maithreyi Raghu, Jascha Sohl-Dickstein, and Surya Ganguli. Exponential expressivity in deep neural networks through transient chaos. In *Advances In Neural Information Processing Systems*, pages 3360–3368, 2016.

[40] Samuel S. Schoenholz, Justin Gilmer, Surya Ganguli, and Jascha Sohl-Dickstein. Deep Information Propagation. 2017. URL https://openreview.net/pdf?id=H1W1UN9gg.

[41] Roland Speicher. Free Probability Theory. *arXiv:0911.0087 [math]*, October 2009. URL http://arxiv.org/abs/0911.0087.

[42] Terence Tao. Topics in random matrix theory. *Graduate studies in Mathematics*, 132, 2012.

[43] Wojciech Tarnowski, Piotr Warchoł, Stanisław Jastrzębski, Jacek Tabor, and Maciej A. Nowak. Dynamical Isometry is Achieved in Residual Networks in a Universal Way for any Activation Function. *arXiv:1809.08848 [cs, stat]*, September 2018. URL http://arxiv.org/abs/1809.08848.

[44] Felix von Oppen, Bertrand I. Halperin, and Ady Stern. Conductivity tensor of striped quantum Hall phases. *Physical Review Letters*, 84(13):2937–2940, March 2000. ISSN 0031-9007, 1079-7114. doi: 10.1103/PhysRevLett.84.2937. URL http://arxiv.org/abs/cond-mat/9910132. arXiv: cond-mat/9910132.

[45] Eugene P. Wigner. Characteristic Vectors of Bordered Matrices With Infinite Dimensions. *The Annals of Mathematics*, 62(3):548, November 1955. ISSN 0003486X. doi: 10.2307/1970079. URL https://www.jstor.org/stable/1970079?origin=crossref.

[46] Eugene P. Wigner. On the Distribution of the Roots of Certain Symmetric Matrices. *The Annals of Mathematics*, 67(2):325, March 1958. ISSN 0003486X. doi: 10.2307/1970008. URL https://www.jstor.org/stable/1970008?origin=crossref.

[47] Christopher K I Williams. Computing with Infinite Networks. In *Advances in neural information processing systems*, page 7, 1997.

[48] Lechao Xiao, Yasaman Bahri, Jascha Sohl-Dickstein, Samuel Schoenholz, and Jeffrey Pennington. Dynamical Isometry and a Mean Field Theory of CNNs: How to Train 10,000-Layer Vanilla Convolutional Neural Networks. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 5393–5402, Stockholmsmässan, Stockholm Sweden, July 2018. PMLR. URL http://proceedings.mlr.press/v80/xiao18a.html.

[49] Greg Yang. Scaling Limits of Wide Neural Networks with Weight Sharing: Gaussian Process Behavior, Gradient Independence, and Neural Tangent Kernel Derivation. *arXiv:1902.04760 [cond-mat, physics:math-ph, stat]*, February 2019.

[50] Greg Yang. Tensor programs i: Wide feedforward or recurrent neural networks of any architecture are gaussian processes. In *Advances in Neural Information Processing Systems*, pages 9947–9960, 2019.

[51] Greg Yang. Tensor programs ii: Neural tangent kernel for any architecture, 2020.

[52] Greg Yang and Sam S. Schoenholz. Deep mean field theory: Layerwise variance and width variation as methods to control gradient explosion, 2018. URL https://openreview.net/forum?id=rJGY8GbR-.

[53] Greg Yang and Samuel S. Schoenholz. Mean Field Residual Network: On the Edge of Chaos. In *Advances in neural information processing systems*, 2017.

[54] Greg Yang, Jeffrey Pennington, Vinay Rao, Jascha Sohl-Dickstein, and Samuel S. Schoenholz. A Mean Field Theory of Batch Normalization. *arXiv:1902.08129 [cond-mat]*, February 2019. URL http://arxiv.org/abs/1902.08129.

[55] Difan Zou, Yuan Cao, Dongruo Zhou, and Quanquan Gu. Stochastic Gradient Descent Optimizes Over-parameterized Deep ReLU Networks. *arXiv:1811.08888 [cs, math, stat]*, November 2018. URL http://arxiv.org/abs/1811.08888.

# A NETSOR$\top$ Master Theorem for Convergence in Mean and in Distribution

## A.1 Convergence in Mean

**Theorem A.1.** *For the same premise as in Theorem 2.10, if $\psi$ is bounded, then we have the convergence in mean*

$$\frac{1}{n}\sum_{\alpha=1}^{n}\psi(h_\alpha^1,\ldots,h_\alpha^k)\xrightarrow{L^1}\mathbb{E}\,\psi(Z^{h^1},\ldots,Z^{h^k}). \tag{16}$$

*This in particular means the convergence of expectations*

$$\frac{1}{n}\mathbb{E}\sum_{\alpha=1}^{n}\psi(h_\alpha^1,\ldots,h_\alpha^k)=\mathbb{E}\,\psi(h_\beta^1,\ldots,h_\beta^k)\to\mathbb{E}\,\psi(Z^{h^1},\ldots,Z^{h^k}),\quad \textit{for any } \beta\geq 1. \tag{17}$$

*In fact, we have almost sure convergence and convergence in mean of* conditional *expectations as well: For each $n$, let $\mathcal{B}$ be a sub-$\sigma$-algebra of the $\sigma$-algebra induced by random sampling in Setup 2.2. Then*

$$\left|\frac{1}{n}\mathbb{E}\left[\sum_{\alpha=1}^{n}\psi(h_\alpha^1,\ldots,h_\alpha^k)\mid\mathcal{B}\right]-\mathbb{E}\,\psi(Z^{h^1},\ldots,Z^{h^k})\right|\xrightarrow{\text{a.s.}}0,\xrightarrow{L^1}0. \tag{18}$$

*Proof.* Eqs. (16) and (18) follow from Theorem 2.10, the boundedness of $\psi$, and (conditional) dominated convergence. The first equality of Eq. (17) follows from the symmetry in $\alpha$, and the convergence of expectations follows from convergence in mean. $\square$

In Theorem A.2 below, we also show convergence in mean when $\psi$ is quadratically bounded and all nonlinearities are linearly bounded. We say a function $\phi:\mathbb{R}^k\to\mathbb{R}$ is *linearly bounded* (resp. *quadratically bounded*) if for all $x_1,\ldots,x_k\in\mathbb{R}$, $\phi(x_1,\ldots,x_k)\leq C(1+|x_1|+\cdots+|x_k|)$ (resp. $C(1+|x_1|^2+\cdots+|x_k|^2)$) for some constant $C>0$. Note that such a linearly bounded $\phi$, applied coordinatewise to vectors $h^1,\ldots,h^k\in\mathbb{R}^n$, preserves $\ell_2$ norm:

$$\begin{aligned}\frac{1}{n}\|\phi(h^1,\ldots,h^k)\|^2&\leq\frac{C^2}{n}\sum_{\alpha=1}^{n}(1+|h_\alpha^1|+\cdots+|h_\alpha^k|)^2\leq\frac{(k+1)C^2}{n}\sum_{\alpha=1}^{n}1+|h_\alpha^1|^2+\cdots+|h_\alpha^k|^2\\&\leq(k+1)C^2\left(1+\frac{\|h^1\|^2}{n}+\cdots+\frac{\|h^k\|^2}{n}\right).\end{aligned} \tag{19}$$

**Theorem A.2.** *Theorem A.1 also holds if $\psi$ is quadratically bounded and all nonlinearities $\phi$ used in* `Nonlin` *are linearly bounded.*

*Proof.* As in the proof of Theorem A.1, it suffices to prove the convergence in mean. Below, we say "random variable $R$ is bounded with high probability" if there exist absolute constants $C,c>0$ such that, for all $r\geq C$, we have $R\leq r$ with probability at least $1-C\exp(-cn)$.

Because each Gaussian matrix $W\in\mathcal{W}$ has bounded operator norm with high probability (Fact A.3), each application of `MatMul` preserves $\ell_2$ norm with high probability. Likewise, each application of `Nonlin` preserves $\ell_2$ norm by Eq. (19). Finally, by classic concentration of measure, each $v\in\mathcal{V}$ has bounded $\ell_2$ norm with high probability. Thus, by induction, all vectors in the program have bounded $\ell_2$ norm with high probability. Because $\psi$ is quadratically bounded, this implies

$$Q\overset{\text{def}}{=}\frac{1}{n}\sum_{\alpha=1}^{n}\psi(h_\alpha^1,\ldots,h_\alpha^k)$$

is bounded with high probability. In particular, this quantity has a subexponential tail

$$\Pr(Q>r)\leq Ce^{-crn},\forall r>C\implies\mathbb{E}[Q\mathbb{I}(Q>C)]\leq\frac{C}{cn}e^{-cCn}\to 0.$$

Thus we may apply standard truncation technique to $Q$, decomposing it as $Q=Q\mathbb{I}(Q\leq C)+Q\mathbb{I}(Q>C)$. The latter converges in mean to 0, as shown above, while the former converges in mean by dominated convergence. $\square$

**Fact A.3** (Upper tail estimate for iid random matrix, Corollary 2.3.5 of Tao [42])**.** *For $W \in \mathbb{R}^{n\times n}$ with $W_{\alpha\beta} \sim \mathcal{N}(0, 1/n)$, there exist absolute constants $C, c > 0$ such that*

$$\Pr(\|W\|_{op} > r) \le Ce^{-crn}$$

*for all $r \ge C$.*

We believe

**Conjecture A.4.** *Theorem A.1 holds for any polynomially bounded $\psi$.*

As in the proof of Theorem A.2, this amounts to proving a tail bound on the LHS of Eq. (16). Intuitively, because the source of the randomness is from Gaussian sampling and the nonlinearities are all polynomially bounded, this quantity should have a sub-Weibull tail beyond some constant upper bound. But making this rigorous is subtle, and we leave this for future work.

### A.2 Convergence in Distribution of Coordinates

**Theorem A.5** (Convergence in Distribution of Coordinates)**.** *Assume the same premise as in Theorem 2.10. For any $\alpha \ge 1$, we have the convergence in distribution of*

$$(h^1_\alpha, \ldots, h^k_\alpha) \xrightarrow{\text{d}} (Z^{h^1}, \ldots, Z^{h^k}).$$

*Proof.* For any bounded continuous function $\psi$, $\mathbb{E}\,\psi(h^1_\alpha, \ldots, h^k_\alpha) \to \mathbb{E}\,\psi(Z^{h^1}, \ldots, Z^{h^k})$ by Theorem A.1. ☐

A slightly more involved version of this symmetry argument also shows that different coordinate slices are independent from one another in the large $n$ limit.

**Theorem A.6.** *Assume the same premise as in Theorem 2.10. For any $\alpha_1, \ldots, \alpha_r \ge 1$, all different for each other, we have the convergence in distribution of*

$$\begin{pmatrix} h^1_{\alpha_1} & \cdots & h^k_{\alpha_1} \\ \vdots & \ddots & \vdots \\ h^1_{\alpha_r} & \cdots & h^k_{\alpha_r} \end{pmatrix} \xrightarrow{\text{d}} \begin{pmatrix} Z^{h^1}_1 & \cdots & Z^{h^k}_1 \\ \vdots & \ddots & \vdots \\ Z^{h^1}_r & \cdots & Z^{h^k}_r \end{pmatrix},$$

*where $(Z^{h^1}_1, \ldots, Z^{h^k}_1), \ldots, (Z^{h^1}_r, \ldots, Z^{h^k}_r)$ are iid copies of $(Z^{h^1}, \ldots, Z^{h^k})$.*

*Proof.* We proceed by induction on $r$. The base case of $r = 1$ has already been proven in Theorem A.5. Now assume the inductive hypothesis is proven for $r - 1$ and we seek to show it is also true for $r$. It suffices to prove, for all bounded continuous functions $f_1, \ldots, f_r : \mathbb{R}^k \to \mathbb{R}$, [31]

$$\mathbb{E} \prod_{i=1}^r f_i(h^1_{\alpha_i}, \ldots, h^k_{\alpha_i}) \to \prod_{i=1}^r \mathbb{E}\, f_i(Z^{h^1}, \ldots, Z^{h^k}).$$

By Theorem 2.10, we have, for all $i \in [r]$,

$$\frac{1}{n} \sum_{\beta=1}^n f_i(h^1_\beta, \ldots, h^k_\beta) \xrightarrow{\text{a.s.}} \mathbb{E}\, f_i(Z^{h^1}, \ldots, Z^{h^k}).$$

Thus, taking the product over all $i \in [r]$ and then taking expectation, we have

$$\frac{1}{n^r} \sum_{\beta_1, \ldots, \beta_r} \mathbb{E} \prod_{i=1}^r f_i(h^1_{\beta_i}, \ldots, h^k_{\beta_i}) \to \prod_{i=1}^r \mathbb{E}\, f_i(Z^{h^1}, \ldots, Z^{h^k}),$$

by dominated convergence and the boundedness of $f_i$. Now in the LHS, only $o(n^r)$ of the summands have some pair from $\beta_1, \ldots, \beta_r$ equal to each other. Each such $o(n^r)$ summands, by induction

[31] or we can just consider the families $f_i(x) = e^{it_i x}$ (or their real and imaginary parts), where $\{t_i\}_{i=1}^r$ varies over $\mathbb{R}^r$.

hypothesis, converges to some finite quantity. Thus their total contribution to the LHS is vanishing with $n$. Hence, we have

$$\mathop{\mathbb{E}}_{\substack{\beta_1,\ldots,\beta_r \\ \text{all distinct}}} \mathbb{E} \prod_{i=1}^{r} f_i(h^1_{\beta_i},\ldots,h^k_{\beta_i}) \xrightarrow{\text{a.s.}} \prod_{i=1}^{r} \mathbb{E}\, f_i(Z^{h^1},\ldots,Z^{h^k}).$$

Finally, we note that the inner expectation in the LHS here is symmetric in all such distinct $\beta_1,\ldots,\beta_r$, so

$$\mathbb{E} \prod_{i=1}^{r} f_i(h^1_{\beta_i},\ldots,h^k_{\beta_i}) \xrightarrow{\text{a.s.}} \prod_{i=1}^{r} \mathbb{E}\, f_i(Z^{h^1},\ldots,Z^{h^k})$$

for any distinct $\beta_1,\ldots,\beta_r$, and in particular, for $\beta_1,\ldots,\beta_r = \alpha_1,\ldots,\alpha_r$, as desired. ☐

### A.3 Extensions to NETSOR$\top^+$ and Programs with Variable Dimensions

All of the theorems above hold as stated for programs with variable dimensions (Appendix F), if these programs are setup as in Setup F.5. Likewise, Theorems A.1, A.5 and A.6 hold for NETSOR$\top^+$ programs (Appendix E) as well, if rank stability (Assumption E.7) is satisfied and all nonlinearities $\varphi^u(-;-)$ are parameter-controlled at $\mathring{\Theta}^u$ (see Theorems E.11 and F.10). Similarly,

**Theorem A.7.** *Theorem A.2 holds for* NETSOR$\top^+$ *programs if rank stability is satisfied and the following conditions on nonlinearities hold:*

- *for every vector $h$ with defining nonlinearity $\varphi^h(-;-)$ and limit parameters $\mathring{\Theta}^h$, $\varphi^h(\vec{x};\mathring{\Theta}^h)$ is linearly bounded in $\vec{x}$, and for any $\Theta$,*

$$|\varphi^h(\vec{x};\Theta) - \varphi^h(\vec{x};\mathring{\Theta}^h)| \leq f(\Theta - \mathring{\Theta}^h)\phi(\vec{x}) \tag{20}$$

  *for some linearly bounded $\phi$ and some continuous function $f$, taking values in $\mathbb{R}$, with $f(0) = 0$.*

- *for every scalar $c$ with defining nonlinearity $\varphi^c(-;-)$ and limit parameters $\mathring{\Theta}^c$, $\varphi^c(\vec{x};\mathring{\Theta}^c)$ is quadratically bounded in $\vec{x}$ and Eq. (20) is satisfied for some quadratically bounded $\phi$ and some continuous function $f$, taking values in $\mathbb{R}$, with $f(0) = 0$.*

# B Generalized Architectural Universality of Neural Network-Gaussian Process

Wide, randomly initialized neural networks are distributed like Gaussian processes (GP) [13, 23, 28, 31, 32, 50]. Yang [50] systematically generalized this result from toy neural networks to all modern neural networks. This result was proved by 1) showing that the kernel of input embeddings converge almost surely to a deterministic kernel $K$, and 2) because the readout layer(s) are independently initialized from the rest of the network, the distribution of random neural network function converges to a GP with this kernel $K$. Part 1 assumed that no weight matrix is the transpose of another weight matrix, and with this assumption, it follows from the NETSOR Master Theorem [50]. However, with the NETSOR$\top$ Master Theorem, we can straightforwardly get rid of this assumption, and the logic above still holds. We thus conclude

**Theorem B.1.** *Consider any neural network whose forward propagation is expressible in a* NETSOR$\top$ *program, and whose output layer(s) are independently sampled from other parameters and not used in the interior of the network. Then we have the convergence in distribution of the neural network function to a Gaussian Process, as the network widths tend to infinity.*[32]

We can also consider what happens when the output layer(s) are not independent from other parameters. For example, we can consider the case where the output can be expressed as the first coordinate $y_1$ of a vector $y = Wu \in \mathbb{R}^n$ for some embedding $u$ of the input. Here $W$ could have been used in the interior of the network. Then in general, $\dot{Z}^y$ is nonzero, and $y_1 \xrightarrow{\text{d}} Z^y$ converges to a non-Gaussian distribution, which can still nevertheless be calculated using Box 1 and Theorem 2.10.

[32]Of course, to compute the GP kernel requires going through the NETSOR$\top$ Master Theorem (in particular, computing $\dot{Z}$ and $\hat{Z}$), and it would be incorrect to just assume all instances of $W^\top$ to be independent from $W$.

## C Generalized Architectural Universality of Neural Tangent Kernel

Jacot et al. [19] showed that, in the limit of large width, a neural network undergoing training by gradient descent evolves like a linear model with a kernel, called the *Neural Tangent Kernel (NTK)*. Yang [51] showed how to calculate this the infinite-width limit of this NTK for any architecture, when a commonly satisfied condition, called *Simple GIA Check*, is satisfied. This condition allows us to assume $W^\top$ to be independent from $W$ in the computation of NTK (this heuristic is called the *Gradient Independence Assumption*, or *GIA*).

*Condition* 1 (Simple GIA Check). No weight matrix is the transpose of another weight matrix; the output layer is sampled independently and with zero mean from all other parameters and is not used anywhere else in the interior of the network[33].

Using Theorem 2.10, we may now generalize the result of Yang [51] to when Simple GIA Check is not satisfied (Theorem C.1). Before that, let's gather some concrete intuition for this generalization.

**Example** Yang [51] demonstrated a simple neural network not satisfying Simple GIA Check, and, if we assume $W^\top$ to be independent from $W$, then the resulting NTK computation would be wrong. Now, with the help of NETSOR$\top$ Master Theorem, we may finally perform the correct computation.

The neural network in question computes

$$x^1 = W^1\xi + 1, \quad h^2 = W^2 x^1, \quad x^2 = \phi(h^2), \quad y = 1^\top x^2/n \tag{21}$$

with $\phi(z) = z^2$ being the square function, $\xi = 0 \in \mathbb{R}^d, y \in \mathbb{R}, x^1, h^2, x^2 \in \mathbb{R}^n, W^1 \in \mathbb{R}^{n\times d}, W^2 \in \mathbb{R}^{n\times n}, W^1_{\alpha\beta} \sim \mathcal{N}(0, 1/d), W^2_{\alpha\beta} \sim \mathcal{N}(0, 1/n)$. If we set $dx^2 = n\frac{\partial y}{\partial x^2}$, then backprop yields

$$dx^2 = 1, \quad dh^2 = 2h^2 \odot 1 = 2h^2, \quad dx^1 = W^{2\top} dh^2 = 2W^{2\top}h^2 = 2W^{2\top}W^2x^1. \tag{22}$$

Yang [51] showed that $\mathbb{E}\, dx^1_\alpha = 2\,\mathbb{E}\, x^1_\alpha = 1$ but if we were to assume $W^{2\top}$ be independent from $W^2$, then we would get the erroneous answer $\mathbb{E}\, dx^1_\alpha = 0$. Here, let us compute $Z^\bullet$ for each vector above using Box 1, and see we get a consistent result with $\mathbb{E}\, Z^{dx^1_\alpha} = 2$.

In the forward pass, we have $Z^{h^2} = \mathcal{N}(0, 1)$ and likewise in the backward pass, $Z^{dh^2} = 2Z^{h^2} = \mathcal{N}(0, 4)$. Next, we need to compute $\hat{Z}^{dx^1}$ and $\dot{Z}^{dx^1}$. Just like in the Master Theorem of Yang [51], $\hat{Z}^{dx^1}$ is the random variable that we would get if we assume $W^\top$ be independent from $W$: $\hat{Z}^{dx^1} = \mathcal{N}(0, \mathbb{E}(Z^{dh^2})^2) = \mathcal{N}(0, 4)$ and is independent from $Z^{dh^2}$ and $Z^{h^2}$. On the other hand, $\dot{Z}^{dx^1}$ is a scalar multiple $\dot{Z}^{dx^1} = x^1\,\mathbb{E}\,\frac{\partial Z^{dh^2}}{\partial h^2}$ of $x^1$. The multiple is $\mathbb{E}\,\frac{\partial Z^{dh^2}}{\partial h^2} = \mathbb{E}\, 2 = 2$. Putting them all together, we get

$$Z^{dx^1} = \hat{Z}^{dx^1} + \dot{Z}^{dx^1} = \mathcal{N}(0, 4) + 2x^1$$

which has mean $\mathbb{E}\, Z^{dx^1} = 2\,\mathbb{E}\, Z^{x^1} = 2$.

More generally, by Theorem 2.10, we trivially have the following result:

**Theorem C.1.** *Consider any neural network whose forward and backpropagation are expressible in a* NETSOR$\top$ *program*[34]*, and which does* not *need to satisfy Simple GIA Check. Suppose the network is parametrized in the* NTK parametrization. *If all of the network's nonlinearities have polynomially bounded weak derivatives, then its NTK, on standard Gaussian initialization of the network parameters, converges to a deterministic kernel as its widths tend to infinity, over any finite set of inputs.*

As with the NNGP in Appendix B, this infinite-width NTK can be computed in a straightforward way using Theorem 2.10, following the examples of Yang [51].

## D Simple GIA Check Implies Gradient Independence Assumption

We give a new proof of the following in this section through FIP.

[33] i.e. if the output weight is $v$ and the output is $v^\top x$, then $x$ does not depend on $v$.

[34] as shown in Yang [51], this includes practically all architectures used in modern deep learning

**Theorem D.1.** *If a neural network expressible*[35] *in a* NETSOR$\top$ *program satisfies Simple GIA Check (Condition 1), then its NTK can be computed assuming that $W^\top$ used in backpropagation is independent from $W$ used in forward propagation.*

Yang [51] showed that, for any pair of inputs $\xi, \bar{\xi} \in \mathbb{R}^d$, the Neural Tangent Kernel of a NN $f$ depends on backpropagation only through the quantities $\langle \nabla_{y(\xi)} f(\xi), \nabla_{y(\bar{\xi})} f(\bar{\xi}) \rangle$. If the network satisfied Simple GIA Check (Condition 1), i.e. the network output is computed like $f(\xi) = n^{-1/2} v^\top e(\xi)$ for some embedding $e(\xi)$ of $\xi$ and $v_\alpha \sim \mathcal{N}(0,1)$ independent of $e$, then we can rewrite $\nabla_{y(\xi)} f(\xi) = n^{-1/2} v^\top J(\xi)$ where $J(\xi) = \partial e(\xi)/\partial y(\xi)$. Therefore,

$$\langle \nabla_{y(\xi)} f(\xi), \nabla_{y(\bar{\xi})} f(\bar{\xi}) \rangle = n^{-1} v^\top J(\xi) J(\bar{\xi})^\top v.$$

By reversing the trace trick (Eq. (6)), this quantity has the same limit as $n^{-1} \operatorname{tr} J(\xi) J(\bar{\xi})^\top$. If $\pi$ is the program expressing the forward and backward propagations of $f(\xi)$ and $f(\bar{\xi})$, then this is a moment in $\mathcal{D}(\pi)$ and $\{W, W^\top\}, W \in \mathcal{W}$. By FIP (Theorem 4.2), this moment stays the same if all $W \in \mathcal{W}$ are assumed independent from $\mathcal{D}(\pi)$, i.e. if transposed weight matrices $W^\top$ in the backward pass are independent from $W$ used in the forward pass (which produced the diagonals of $\mathcal{D}(\pi)$).

# E NETSOR$\top^+$: Adding Scalars to NETSOR$\top$

In this section, we discuss the extension of NETSOR$\top$ with scalars that can be computed from (essentially) averaging some vector in the program. This is not used substantially until Appendix I, so the reader should feel free to skip ahead and come back only as needed.

**Definition E.1.** A NETSOR$\top^+$ program[36] is just a sequence of $\mathbb{R}^n$ vectors and scalars in $\mathbb{R}$ inductively generated via one of the following ways from an initial set $\mathcal{C}$ of random scalars, an initial set $\mathcal{V}$ of random $\mathbb{R}^n$ vectors, and a set $\mathcal{W}$ of random $\mathbb{R}^{n\times n}$ matrices

`Nonlin`$^+$ Given $\phi : \mathbb{R}^k \times \mathbb{R}^l \to \mathbb{R}$, previous scalars $\theta_1, \dots, \theta_l \in \mathbb{R}$ and vectors $x^1, \dots, x^k \in \mathbb{R}^n$, we can generate a new vector

$$\phi(x^1, \dots, x^k; \theta_1, \dots, \theta_l) \in \mathbb{R}^n$$

where $\phi(-; \theta_1, \dots, \theta_l)$ applies coordinatewise to each "$\alpha$-slice" $(x^1_\alpha, \dots, x^k_\alpha)$.

`Moment` Given same setup as above, we can also generate a new scalar

$$\frac{1}{n} \sum_{\alpha=1}^{n} \phi(x^1_\alpha, \dots, x^k_\alpha; \theta_1, \dots, \theta_l) \in \mathbb{R}.$$

`MatMul` Given $W \in \mathbb{R}^{n\times n}$ and $x \in \mathbb{R}^n$, we can generate $Wx \in \mathbb{R}^n$ or $W^\top x \in \mathbb{R}^n$

We say a vector is a *G-var* if it is in $\mathcal{V}$ or it is introduced by `MatMul`.

We will typically discuss NETSOR$\top^+$ programs when $\mathcal{C}$, $\mathcal{V}$, and $\mathcal{W}$ are sampled as follows. Note $\mathcal{V}$ and $\mathcal{W}$ are sampled the same way here as in Setup 2.2.

**Setup E.2** (NETSOR$\top^+$)**.** *1) each random scalar $c$ in $\mathcal{C}$ converges to a deterministic limit $\mathring{c} \in \mathbb{R}$ as $n \to \infty$; 2) for each initial $W \in \mathcal{W}$, $W_{\alpha\beta} \sim \mathcal{N}(0, \sigma_W^2/n)$ for an associated variance $\sigma_W^2$; 3) there is a multivariate Gaussian $Z^{\mathcal{V}} = \{Z^g : g \in \mathcal{V}\} \in \mathbb{R}^{|\mathcal{V}|}$ such that the initial set of vectors $\mathcal{V}$ are sampled like $\{g_\alpha : g \in \mathcal{V}\} \sim Z^{\mathcal{V}}$ iid for each $\alpha \in [n]$.*

**Definition E.3** (Key Intuition for NETSOR$\top^+$)**.** Just like in the NETSOR$\top$ case, each vector $h$ in the program has roughly iid coordinates when $n \gg 1$, each of which is distributed like a random variable $Z^h$. In addition, each scalar $\theta$ in the program will converge to a deterministic limit $\mathring{\theta}$. We'll recursively define $Z^h$ and $\mathring{\theta}$ as follows.

[35] i.e. its forward and backward propagations at initialization can be written in a NETSOR$\top$ program. As shown in Yang [50, 51], this covers almost all classic and modern neural networks.

[36] What we refer to as NETSOR$\top^+$ program is the same as "simplified NETSOR$\top^+$" in Yang [51]

`ZInit` If $h \in \mathcal{V}$, then $Z^h$ is defined as the distribution of each coordinate of $h$ given in Setup E.2. We also set $\hat{Z}^h \stackrel{\text{def}}{=} Z^h$ and $\dot{Z}^h \stackrel{\text{def}}{=} 0$. Likewise, if $\theta \in \mathcal{C}$, then $\mathring{\theta}$ is defined as the limit of $\theta$ specified in Setup E.2.

`ZMatMul` Same as in `ZMatMul` in NETSOR$\top$

`ZNonlin`$^+$ Given $\phi : \mathbb{R}^k \times \mathbb{R}^l \to \mathbb{R}$, previous scalars $\theta_1, \ldots, \theta_l \in \mathbb{R}$ and vectors $x^1, \ldots, x^k \in \mathbb{R}^n$, we have

$$Z^{\phi(x^1,\ldots,x^k;\theta_1,\ldots,\theta_l)} \stackrel{\text{def}}{=} \phi(Z^{x^1}, \ldots, Z^{x^k}; \mathring{\theta}_1, \ldots, \mathring{\theta}_l).$$

`ZMoment` Given same setup as above and scalar $\theta = \frac{1}{n}\sum_{\alpha=1}^n \phi(x^1_\alpha, \ldots, x^k_\alpha; \theta_1, \ldots, \theta_l)$, then

$$\mathring{\theta} \stackrel{\text{def}}{=} \mathbb{E}\,\phi(Z^{x^1}, \ldots, Z^{x^k}; \mathring{\theta}_1, \ldots, \mathring{\theta}_l).$$

Here $\mathring{\theta}_1, \ldots, \mathring{\theta}_l$ are deterministic, so the expectation is taken over $Z^{x^1}, \ldots, Z^{x^k}$.

### E.1 NETSOR$\top^+$ Master Theorem for Pseudo-Lipschitz Nonlinearities

**Pseudo-Lipschitz functions** are, roughly speaking, functions whose weak derivatives are polynomially bounded.

**Definition E.4.** A function $f : \mathbb{R}^k \to \mathbb{R}$ is called *pseudo-Lipschitz* of degree $d$ if $|f(x) - f(y)| \le C\|x - y\|(1 + \sum_{i=1}^k |x_i|^d + |y_i|^d)$ for some $C$.

Here are some basic properties of pseudo-Lipschitz functions:

- The norm $\|\cdot\|$ in Definition E.4 can be any norm equivalent to the $\ell_2$ norm, e.g. $\ell_p, p \ge 1$, norms. Similarly, $\sum_{i=1}^k |x_i|^d + |y_i|^d$ can be replaced by $\|x\|_p^d + \|y\|_p^d$, for any $p \ge 1$.
- A pseudo-Lipschitz function is polynomially bounded.
- A composition of pseudo-Lipschitz functions of degrees $d_1$ and $d_2$ is pseudo-Lipschitz of degree $d_1 + d_2$.
- A pseudo-Lipschitz function is Lipschitz on any compact set.

**Master Theorem** We state the Master Theorem below assuming a generic regularity condition called *rank stability* (Assumption E.7), which we shall describe shortly. The proof will follow from a more general, but more wordy Master Theorem in the next section (Theorem E.11).

**Theorem E.5** (Pseudo-Lipschitz NETSOR$\top^+$ Master Theorem)**.** *Fix a* NETSOR$\top^+$ *program. Suppose the initial matrices $\mathcal{W}$, vectors $\mathcal{V}$, and scalars $\mathcal{C}$ are sampled in the fashion of Setup E.2. Suppose the program satisfies the* rank stability assumption *below (Assumption E.7). Assume all $\phi$ used in* `Nonlin`$^+$ *are pseudo-Lipschitz (or, we can assume the slightly weaker Assumption E.6). Then*

- *For any fixed $k$ and any polynomially-bounded $\psi : \mathbb{R}^k \to \mathbb{R}$, as $n \to \infty$,*

$$\frac{1}{n}\sum_{\alpha=1}^n \psi(h^1_\alpha, \ldots, h^k_\alpha) \xrightarrow{\text{a.s.}} \mathbb{E}\,\psi(Z^{h^1}, \ldots, Z^{h^k}),$$

  *for any vectors $h^1, \ldots, h^k$ in the program, where $Z^{h^i}$ are as defined in Definition E.3.*

- *Any scalar $\theta$ in the program tends to $\mathring{\theta}$ almost surely, where $\mathring{\theta}$ is as defined in Definition E.3.*

**Assumption E.6.** *Suppose*

1. *If a function $\phi(;-) : \mathbb{R}^{0+l} \to \mathbb{R}$ with only parameter arguments is used in* `Moment`*, then $\phi$ is continuous in those arguments.*
2. *Any other function $\phi(-;-) : \mathbb{R}^{k+l} \to \mathbb{R}$ with parameters (where $k > 0$) used in* `Nonlin`$^+$ *or* `Moment` *is pseudo-Lipschitz in all of its arguments (both inputs and parameters).*

Statement 1 in Assumption E.6 essentially says that if we have scalars $\theta_1, \ldots, \theta_l$ in the program, then we can produce a new scalar by applying a continuous function (a weaker restriction than a pseudo-Lipschitz function) to them. Indeed, if $\theta_1, \ldots, \theta_l$ converge almost surely, then this new scalar does too.

**Rank Stability** The following assumption says that the vectors in a program should not change any linear dependence relations abruptly in the infinite $n$ limit.

**Assumption E.7** (Rank Stability)**.** *Fix a* NETSOR$\top^+$ *program that is setup by Setup E.2. We say this program satisfies* rank stability *if for any matrix $W \in \mathbb{R}^{n\times n}$ in the program*[37] *and any collection $\mathcal{H}$ of vectors $h$ such that $Wh$ appears in the program, we have* $\operatorname{rank}\mathcal{H} = \operatorname{rank}\{Z^h : h \in \mathcal{H}\}$*, almost surely, for sufficiently large $n$.*[38]

Most commonly, $\{Z^h : h \in \mathcal{H}\}$ will be linearly independent for all such $\mathcal{H}$, and therefore, by the lower semicontinuity of rank, Assumption E.7 is automatically satisfied. We shall discuss the necessity of Assumption E.7 below (Remark E.13).

### E.2 NETSOR$\top^+$ Master Theorem for Parameter-Controlled Nonlinearities

**Parameter-Control** Theorem E.5 will follow from the more general Master Theorem we state in this section, which allows for more general nonlinearities which only needs to be *mildly* smooth in the scalar parameters $\theta_1, \ldots, \theta_l$, but not necessarily in $x^1_\alpha, \ldots, x^k_\alpha$:

**Definition E.8** (Parameter-Control)**.** We say a parametrized function $\phi(-;-) : \mathbb{R}^k \times \mathbb{R}^l \to \mathbb{R}$ is *polynomially parameter-controlled* or just *parameter-controlled* for short[39] , at $\mathring{\Theta} \in \mathbb{R}^l$ if

1. $\phi(-;\mathring{\Theta})$ is polynomially bounded, and
2. there are some polynomially bounded $\bar{\phi} : \mathbb{R}^k \to \mathbb{R}$ and some function $f : \mathbb{R}^l \to \mathbb{R}^{\geq 0} \cup \{\infty\}$ that has $f(\mathring{\Theta}) = 0$ and that is continuous at $\mathring{\Theta}$, such that, for all $x^1, \ldots, x^k \in \mathbb{R}$ and $\Theta \in \mathbb{R}^l$,
$$|\phi(x^1, \ldots, x^k; \Theta) - \phi(x^1, \ldots, x^k; \mathring{\Theta})| \leq f(\Theta)\bar{\phi}(x^1, \ldots, x^k).$$

Note that $f$ and $\bar{\phi}$ here can depend on $\mathring{\Theta}$. The following examples come from Yang [50].

*Example* E.9. Any function that is (pseudo-)Lipschitz in $x^1, \ldots, x^k$ and $\Theta$ is polynomially parameter-controlled. An example of a discontinuous function that is polynomially parameter-controlled is $\phi(x;\theta) = \operatorname{step}(\theta x)$: For $\mathring{\theta} \neq 0$, we have
$$|\phi(x;\theta) - \phi(x;\mathring{\theta})| \leq \frac{|\mathring{\theta} - \theta|}{|\mathring{\theta}|},$$
so we can set $f(\theta) = \frac{|\mathring{\theta}-\theta|}{|\mathring{\theta}|}$ and $\bar{\phi} = 1$ in Definition E.8.

Next, note that we can always express a vector in a NETSOR$\top^+$ program as a nonlinear image of previous G-vars. For example, if $z = \phi(x^1, x^2; \theta_1), x^1 = Wv, x^2 = \psi(y; \theta_2), y = Wu$, then $z$ can be directly expressed in terms of G-vars: $z = \hat{\phi}(Wv, Wu; \theta_1, \theta_2) \stackrel{\text{def}}{=} \phi(Wv, \psi(Wu; \theta_2); \theta_1)$. Therefore, we can make the following definition

**Definition E.10** ($\varphi^\bullet, \Theta^\bullet$ Notation)**.** For any $\mathbb{R}^n$ vector $x$ in a NETSOR$\top^+$ program (Definition E.1), let $\varphi^x$ and $\Theta^x$ be the parametrized nonlinearity and the scalars such that $x = \varphi^x(z^1, \ldots, z^k; \Theta^x)$ for some G-vars $z^1, \ldots, z^k$. Likewise, for any scalar $c$ in a NETSOR$\top^+$ program, let $\varphi^c$ and $\Theta^c$ be the parametrized nonlinearity and the scalars such that $c = \frac{1}{n}\sum_{\alpha=1}^n \varphi^x(z^1_\alpha, \ldots, z^k_\alpha; \Theta^c)$ for some G-vars $z^1, \ldots, z^k$.

**Theorem E.11** (Parameter-Controlled NETSOR$\top^+$ Master Theorem)**.** *Fix a* NETSOR$\top$ *program. Suppose the initial matrices $\mathcal{W}$, vectors $\mathcal{V}$, and scalars $\mathcal{C}$ are sampled in the fashion of Setup E.2. Suppose the program satisfies the* rank stability assumption *(Assumption E.7). Assume $\varphi^u(-;-)$ is parameter-controlled at $\mathring{\Theta}^u$ for all vectors and scalars $u$. Then*

[37] i.e. either $W \in \mathcal{W}$ or $W^\top \in \mathcal{W}$

[38] In the case of variable dimension NETSOR$\top^+$ programs, this is the same except $W \in \mathbb{R}^{n\times m}$ can have unequal dimensions and the limit is taken in the manner of Setup F.5.

[39] This overloads the meaning of *parameter-controlled* from Yang [50], where the definition replaces the "polynomially bounded" in the definition here with "bounded by $e^{C\|\cdot\|^{2-\epsilon}+c}$ for some $C, c, \epsilon > 0$." In this paper, we shall never be concerned with the latter (more generous) notion of boundedness, so there should be no risk of confusion.

- *For any random vector $\Theta \in \mathbb{R}^l$ that converges almost surely to a deterministic vector $\mathring{\Theta}$ as $n \to \infty$, and for any $\psi(-;-) : \mathbb{R}^k \times \mathbb{R}^l \to \mathbb{R}$ parameter-controlled at $\mathring{\Theta}$,*

$$\frac{1}{n}\sum_{\alpha=1}^{n} \psi(g^1_\alpha, \ldots, g^k_\alpha; \Theta) \xrightarrow{\text{a.s.}} \mathbb{E}\, \psi(Z^{g^1}, \ldots, Z^{g^k}; \mathring{\Theta}).$$

  *for any G-vars $g^1, \ldots, g^k$, where $Z^{g^i}$ are as defined in Definition E.3.*

- *Any scalar $\theta$ in the program tends to $\mathring{\theta}$ almost surely, where $\mathring{\theta}$ is as defined in Definition E.3.*

Since pseudo-Lipschitz parametrized functions and parameterless polynomially bounded functions are both parameter-controlled, Theorem E.11 implies Theorem E.5 trivially. Proof of Theorem E.11 can be found in Appendix M

**Necessity of the Master Theorems' Assumptions** The following remarks from [50] show the necessity of parameter control and of rank stability in Theorem E.11.

*Remark* E.12 (Necessity of parameter-control). Suppose $\psi(x; \theta) = \mathbb{I}(\theta x \neq 0)$. For $\theta \neq 0$, $\psi$ is 1 everywhere except $\psi(0; \theta) = 0$. For $\theta = 0$, $\psi$ is identically 0. Thus it's easily seen that $\psi$ is not parameter-controlled at $\theta = 0$.

Now, if $g \in \mathbb{R}^n$ is sampled like $g_\alpha \sim \mathcal{N}(0, 1)$ and if $\theta = 1/n$ so that $\theta \to \mathring{\theta} = 0$, then

$$\frac{1}{n}\sum_{\alpha=1}^{n} \psi(g_\alpha; \theta) \xrightarrow{\text{a.s.}} 1$$

but

$$\mathbb{E}\, \psi(Z^g; \mathring{\theta}) = \mathbb{E}\, 0 = 0.$$

So our Master Theorem can't hold in this case.

*Remark* E.13 (Necessity of Rank Stability Assumption E.7). Suppose we have two Initial G-vars $g^1, g^2 \in \mathbb{R}^n$ which are sampled independently as $g^1_\alpha, g^2_\alpha \sim \mathcal{N}(0, 1)$. Let $W \in \mathbb{R}^{n\times n}$ be sampled as $W_{\alpha\beta} \sim \mathcal{N}(0, 1/n)$. Then we can define $h^2 = \theta g^2$ where $\theta = \exp(-n)$ as a function of $n$, using `Nonlin`$^+$, so that $h^2_\alpha \xrightarrow{\text{a.s.}} 0$. Additionally, let $\bar{g}^1 = W g^1$ and $\bar{g}^2 = W h^2$. Again, $\bar{g}^2_\alpha \xrightarrow{\text{a.s.}} 0$ but for any finite $n$, $\bar{g}^2$ is linearly independent from $\bar{g}^1$. Thus rank stability does not hold here, since the rank of $\{\bar{g}^1, \bar{g}^2\}$ is 2 for all finite $n$, yet $\{Z^{\bar{g}^1}, Z^{\bar{g}^2}\} = \{\mathcal{N}(0, 1), 0\}$ has rank 1.

Now consider the (parameterless) nonlinearity $\psi(x, y)$ that is 1 except on the line $y = 0$, where it is 0. Then

$$\frac{1}{n}\sum_{\alpha=1}^{n} \psi(\bar{g}^1_\alpha, \bar{g}^2_\alpha) \xrightarrow{\text{a.s.}} 1$$

since $\bar{g}^2_\alpha$ is almost surely nonzero, but

$$\mathbb{E}\, \psi(Z^{\bar{g}^1}, Z^{\bar{g}^2}) = \mathbb{E}\, \psi(Z^{\bar{g}^1}, 0) = \mathbb{E}\, 0 = 0.$$

Note this example uses the non-smoothness of $\psi$ in an essential way. Therefore, we conjecture that rank stability assumption can be removed in Theorem E.5.

*Remark* E.14 (Rank stability already holds for NETSOR$\top$ programs). It turns out that, as long as we only have parameterless nonlinearities, we get rank stability Assumption E.7 for free. This is formulated explicitly in Lemma L.11. It is as a result of our proof of Theorem 2.10 that interleaves an inductive proof of this rank stability (more generally, the inductive hypothesis CoreSet) with an inductive proof of the "empirical moment" convergence (the inductive hypothesis Moments).

**Relative Utilities of Theorem E.5 and Theorem E.11** While pseudo-Lipschitz functions are closed under composition (at the expense of increasing degree), this is in general not true for parameter-control. Thus, the assumptions of Theorem E.11 are relatively more cumbersome to check: one has to manually unwind the nonlinearities into $\varphi^x$ and then check this nested function is parameter-controlled. However, this increased flexibility of the parameter-control condition allows us to prove useful theorems that is not possible with the just Theorem E.5. For example, if $\varphi^x(z; \theta) = z/\theta$, then $\varphi^x$ is not pseudo-Lipschitz, but it is parameter-controlled if $\mathring{\theta} > 0$. This fact is used in Yang [50, 51] to compute the Gaussian Process kernel and the Neural Tangent Kernel of a randomly initialized neural network with layernorm.

### E.3 Getting Rid of Rank Stability Assumption through More Test Function Smoothness

Sometimes, the rank stability assumption Assumption E.7 can be difficult to check, so it is convenient to have a version of the Master Theorem without it. As shown in Remark E.13, to do so, it's not enough to only have parameter-control; we need to have some smoothness in the nonlinearities and the test functions. It turns out we can obtain a version of Theorem E.5 without assuming rank stability.

**Theorem E.15** (Pseudo-Lipschitz NETSOR$\top^+$ Master Theorem without Rank Stability)**.** *Fix a Tensor Program initialized accordingly to Setup E.2. Assume all nonlinearities are pseudo-Lipschitz (or, we can assume the slightly weaker Assumption E.6). Then*

1. *For any fixed $k$ and any pseudo-Lipschitz $\psi : \mathbb{R}^k \to \mathbb{R}$, as $n \to \infty$,*
$$\frac{1}{n}\sum_{\alpha=1}^{n} \psi(h^1_\alpha, \dots, h^k_\alpha) \xrightarrow{\text{a.s.}} \mathbb{E}\, \psi(Z^{h^1}, \dots, Z^{h^k}),$$
*for any vectors $h^1, \dots, h^k$ in the program, where $Z^{h^i}$ are as defined in Definition E.3.*
2. *Any scalar $\theta$ in the program tends to $\mathring{\theta}$ almost surely, where $\mathring{\theta}$ is as defined in Definition E.3.*

**Difference between Theorem E.5 and Theorem E.15** is 1) Theorem E.5 needs rank stability Assumption E.7 and 2) the test function $\psi$ in Theorem E.15(1) is pseudo-Lipschitz.

**Proving Theorem E.15** The main difficulty with proving Theorem E.5 without rank stability is that even if the covariance matrix of the $Z$ random variables converges, its pseudo-inverse may not. This is a crucial step in the Gaussian conditioning trick, so we need to more carefully modify the proof skeleton of Theorem 2.10 (in contrast to the proof of Theorem E.11, which just needed to modify some minor bounds in the proof skeleton). The main idea is to carefully extend the core-set argument of Appendix L. See Appendix N for the proof.

## F Programs with Variable Dimension

We can also allow the matrices and vectors to have variable dimensions (not all equal to $n$). This is useful in reasoning about neural networks of varying widths in each layer, and in proving the Marchenko-Pastur law for arbitrary rectangular shape ratio (see Appendix H). The reader should feel free to skip ahead and read this section only when approaching these topics.

Here we need to spend a few words on the right analogue of "large-$n$" limit we should take. The basic idea is that, every time we apply one of the `MatMul`, `Nonlin`$^+$, or `Moment` rules, we implicitly set equal some dimensions of the matrices and/or vectors involved. These equalities partitions the vectors into classes sharing common dimensions. The limit we shall take is one in which the dimension of each class tends to infinity, such that the dimension ratio of each pair of distinct classes tends to a fixed number strictly in $(0, \infty)$.

**Notation** In this section, we let $\dim(x)$ denote the dimension of a vector $x$. We present a relatively self-contained exposition, but inevitably this will be highly similar to the nonvariable-dimension case, so we highlight important differences in red.

**Definition F.1.** A NETSOR$\top$ program (with variable dimensions) is just a sequence of vectors inductively generated via one of the following ways from an initial set $\mathcal{V}$ of random vectors and a set $\mathcal{W}$ of random matrices

`Nonlin` (Same as in Definition 2.1) Given $\phi : \mathbb{R}^k \to \mathbb{R}$ and $x^1, \dots, x^k \in \mathbb{R}^n$ in the same CDC, we can generate $\phi(x^1, \dots, x^k) \in \mathbb{R}^n$

`MatMul` Given $W \in \mathbb{R}^{n\times m}$ (resp. $W \in \mathbb{R}^{m\times n}$) and $x \in \mathbb{R}^m$, we can generate $Wx \in \mathbb{R}^n$ (resp. $W^\top x \in \mathbb{R}^n$)

**Definition F.2.** A NETSOR$\top^+$ program (with variable dimension) is just a sequence of vectors inductively generated via one of the following ways from an initial set $\mathcal{C}$ of random scalars, an initial set $\mathcal{V}$ of random vectors, and a set $\mathcal{W}$ of random matrices

`Nonlin`$^+$ (Same as in Definition E.1) Given $\phi : \mathbb{R}^k \times \mathbb{R}^l \to \mathbb{R}$, previous scalars $\theta_1, \ldots, \theta_l \in \mathbb{R}$ and vectors $x^1, \ldots, x^k \in \mathbb{R}^n$ in the same CDC, we can generate a new vector

$$\phi(x^1, \ldots, x^k; \theta_1, \ldots, \theta_l) \in \mathbb{R}^n$$

where $\phi(-; \theta_1, \ldots, \theta_l)$ applies coordinatewise to each $\alpha$-slice $(x^1_\alpha, \ldots, x^k_\alpha)$.

`Moment` (Same as in Definition E.1) Given same setup as above, we can also generate a new scalar

$$\frac{1}{n}\sum_{\alpha=1}^{n} \phi(x^1_\alpha, \ldots, x^k_\alpha; \theta_1, \ldots, \theta_l) \in \mathbb{R}$$

`MatMul` Same as in Definition F.1 above.

Note that in both definitions above, the essential change is that the matrix in `MatMul` is now allowed to have different dimensions on two sides.

The initial vectors in $\mathcal{V}$ can have varying dimensions. When we take the "large dimension" limit, we may seek to hold certain pairs of such dimensions to be equal. We formalize this as an equivalence relation between vectors in $\mathcal{V}$: $x \simeq y$ if we hold $\dim(x) = \dim(y)$ as this dimension tends to infinity. In addition, we have the following natural constraints on the dimensions of vectors involved in each rule.

$$\begin{cases} \text{If } y = \phi(x^1, \ldots, x^k), \text{ then } \dim(y) = \dim(x^i), \forall i. \\ \text{If } y = Wx \text{ and } \bar{y} = W\bar{x}, \text{ then } \dim(x) = \dim(\bar{x}) \text{ and } \dim(y) = \dim(\bar{y}). \end{cases} \tag{23}$$

This allows us to extend the equivalence relation $\simeq$ on $\mathcal{V}$ as discussed above to all vectors in the program.

**Definition F.3.** Given an equivalence relation $\simeq$ on the vectors of a program, we extend this to an equivalence relation on all vectors of the program as the smallest equivalence relation containing the relation

$$h \equiv h' \iff h \simeq h' \text{ OR } h \text{ and } h' \text{ are constrained to have the same dimension by (23).} \tag{24}$$

We call any such equivalence class a *Common Dimension Class*, or CDC. We denote this common dimension of vectors in a CDC $\mathfrak{c}$ by $\dim(\mathfrak{c})$.

Intuitively, the dimensions of vectors in each CDC are all the same but can be different in different CDCs. The CDCs form a partition of all the vectors in a program. Each matrix in $\mathcal{W}$ "straddles" between two (possible the same) CDCs.

*Example* F.4. Consider the MLP in Eq. (2) but with variable widths: $f(\xi; \theta) = W^{L+1}x^L(\xi)$ with input $\xi \in \mathbb{R}^{n^0}$ and output dimension $n^{L+1} = 1$, where we recursively define, for $l = 2, \ldots, L$,

$$h^l(\xi) = W^l x^{l-1}(\xi) \in \mathbb{R}^{n^l}, \quad x^l(\xi) = \phi(h^l(\xi)), \quad h^1(\xi) = W^1\xi \in \mathbb{R}^{n^1}$$

where each $W^l \in \mathbb{R}^{n^l \times n^{l-1}}$. Suppose we sample $W^l_{\alpha\beta} \sim \mathcal{N}(0, 1/n^{l-1})$. The vectors in the program are $h^1(\xi), x^1(\xi), \ldots, h^l(\xi), x^l(\xi)$. In Definition F.3, if $\simeq$ is empty, then the CDCs will just be $\{h^1(\xi), x^1(\xi)\}, \ldots, \{h^l(\xi), x^l(\xi)\}$ (each with common dimension $n^l$). If instead we set $h^1(\xi) \simeq \cdots \simeq h^l(\xi)$, then there is only one CDC with common dimension $n^1 = \cdots = n^L$.

**Setup F.5** (Variable Dimension NETSOR$\top$ or NETSOR$\top^+$)**.** *We will consider* NETSOR$\top$ *or* NETSOR$\top^+$ *programs initialized as follows.*

**Matrix sampling** *For each initial $W \in \mathbb{R}^{n \times m}$ in $\mathcal{W}$, we sample $W_{\alpha\beta} \sim \mathcal{N}(0, \sigma_W^2/m)$ for an associated variance $\sigma_W^2$. In this setting, we also set $\sigma^2_{W^\top} \overset{\text{def}}{=} \frac{n}{m}\sigma_W^2$ so that $(W^\top)_{\alpha\beta} \overset{\text{d}}{=} \mathcal{N}(0, \sigma^2_{W^\top}/n)$.*

**Vector sampling** *For every CDC $\mathfrak{c}$, there is a multivariate Gaussian $Z^{\mathfrak{c}\cap\mathcal{V}} = \{Z^g : g \in \mathfrak{c} \cap \mathcal{V}\} \in \mathbb{R}^{\mathfrak{c}\cap\mathcal{V}}$ such that the vectors in $\mathfrak{c} \cap \mathcal{V}$ are sampled like $\{g_\alpha : g \in \mathfrak{c} \cap \mathcal{V}\} \sim Z^{\mathfrak{c}\cap\mathcal{V}}$ iid for each $\alpha \in [\dim(\mathfrak{c})]$.*

**Scalar sampling** *For a* NETSOR$\top^+$ *program, the scalars in $\mathcal{C}$ are sampled the same way as in Setup E.2.*

**How the limit is taken** *We shall consider a limit where for every matrix $W \in \mathbb{R}^{n\times m}$ in $\mathcal{W}$, the dimensions $n, m$ (which are dimensions of corresponding CDCs) tend to $\infty$ such that their ratio $n/m \to \rho \in (0,\infty)$ for some finite but nonzero $\rho$. Note this implies $\sigma^2_{W^\top} = \frac{n}{m}\sigma^2_W \to \mathring{\sigma}^2_{W^\top} \stackrel{\text{def}}{=} \rho\sigma^2_W$. We also define $\mathring{\sigma}^2_W \stackrel{\text{def}}{=} \sigma^2_W$ for convenience.*

**Definition F.6.** Given this setup, each vector $h$ again has roughly iid coordinates distributed like $Z^h$, which is defined as follows. Likewise, for NETSOR$\top^+$ programs, each scalar $\theta$ will tend to a deterministic limit $\mathring{\theta}$, as defined below. This is exactly the same as in Box 1 or Definition E.3 except that $\sigma_W$ should be replaced by $\mathring{\sigma}_W$. We highlight the places where this occurs in red below.

`ZInit` Same as in Box 1 or Definition E.3.

`ZMatMul` For every $W_{\alpha\beta} \sim \mathcal{N}(0, \sigma^2_W/n)$ and vector $x$ in the program, we set $Z^{Wx} \stackrel{\text{def}}{=} \hat{Z}^{Wx} + \dot{Z}^{Wx}$ where

`ZHat` $\hat{Z}^{Wx}$ is a Gaussian variable with zero mean. Let $\mathcal{V}_W$ denote the set of all vectors in the program of the form $Wy$ for some $y$. Then $\{\hat{Z}^{Wy} : Wy \in \mathcal{V}_W\}$ is defined to be jointly Gaussian with zero mean and covariance

$$\operatorname{Cov}\left(\hat{Z}^{Wx}, \hat{Z}^{Wy}\right) \stackrel{\text{def}}{=} \mathring{\sigma}^2_W \,\mathbb{E}\, Z^x Z^y, \quad \text{for any } Wx, Wy \in \mathcal{V}_W.$$

Furthermore, $\{\hat{Z}^{Wy} : Wy \in \mathcal{V}_W\}$ is mutually independent from $\{\hat{Z}^v : v \in \mathcal{V} \cup \bigcup_{\bar{W}\neq W} \mathcal{V}_{\bar{W}}\}$, where $\bar{W}$ ranges over $\mathcal{W} \cup \{A^\top : A \in \mathcal{W}\}$.

`ZDot` With partial derivative interpreted naively (but see Remark 2.11 and F.7),

$$\dot{Z}^{Wx} \stackrel{\text{def}}{=} \mathring{\sigma}^2_W \sum Z^y \,\mathbb{E}\, \frac{\partial Z^x}{\partial \hat{Z}^{W^\top y}},$$

summing over $W^\top y \in \mathcal{V}_{W^\top}$ introduced before $x$.

`ZNonlin` Same as in Box 1 or Definition E.3.

`ZMoment` Same as in Definition E.3.

*Remark* F.7 (Partial derivative expectation). The partial derivative expectation $\mathbb{E}\,\frac{\partial Z^x}{\partial \hat{Z}^{W^\top x^i}}$ can be defined without derivatives as in Remark 2.12, with the only change being the dependence on the dimension ratio, which we highlight in red below: In `ZDot` above, suppose $\{W^\top y^i\}_{i=1}^k$ are all elements of $\mathcal{V}_{W^\top}$ introduced before $x$. Define the matrix $C \in \mathbb{R}^{k\times k}$ by $C_{ij} \stackrel{\text{def}}{=} \mathbb{E}\, Z^{y^i} Z^{y^j}$ and define the vector $b \in \mathbb{R}^k$ by $b_i \stackrel{\text{def}}{=} \mathbb{E}\, \hat{Z}^{W^\top y^i} Z^x$. If $a = \rho^{-1} C^+ b$ (where $C^+$ denotes the pseudoinverse of $C$ and $\rho = \lim m/n$ is the limit ratio of dimensions of $W \in \mathbb{R}^{m\times n}$), then in `ZDot` we may set

$$\sigma^2_W \,\mathbb{E}\, \frac{\partial Z^x}{\partial \hat{Z}^{W^\top y^i}} = a_i. \tag{25}$$

Then by straightforward modifications of the proofs of the nonvariable-dimension cases, we can prove the following Master Theorems.

**Theorem F.8** (Variable Dimension NETSOR$\top$ Master Theorem)**.** *Fix a* NETSOR$\top$ *program (with variable dimensions). Suppose the initial matrices $\mathcal{W}$ and vectors $\mathcal{V}$ are sampled in the fashion of Setup F.5. Assume all $\phi$ used in* `Nonlin` *are polynomially bounded. Then for any fixed $k$ and any* polynomially bounded $\psi : \mathbb{R}^k \to \mathbb{R}$, *as the dimensions of the vectors tend to $\infty$ as specified in Setup F.5, we have*

$$\frac{1}{n}\sum_{\alpha=1}^{n} \psi(h^1_\alpha, \ldots, h^k_\alpha) \xrightarrow{\text{a.s.}} \mathbb{E}\,\psi(Z^{h^1}, \ldots, Z^{h^k}),$$

*for any collection of vectors $h^1, \ldots, h^k \in \mathbb{R}^n$ in the program, where $Z^{h^i}$ are defined in Definition F.6.*

**Theorem F.9** (Variable Dimension Pseudo-Lipschitz NETSOR$\top^+$ Master Theorem)**.** *Fix a* NETSOR$\top$ *program (with variable dimensions). Suppose the initial matrices $\mathcal{W}$, vectors $\mathcal{V}$, and scalars $\mathcal{C}$ are sampled in the fashion of Setup F.5. Suppose the program satisfies the* rank stability assumption *(Assumption E.7). Assume each nonlinearity of* `Nonlin`$^+$ *is pseudo-lipschitz. Then*

- *For any fixed $k$ and any polynomially-bounded $\psi : \mathbb{R}^k \to \mathbb{R}$, in the limit described in Setup F.5,*

$$\frac{1}{n}\sum_{\alpha=1}^{n} \psi(h_\alpha^1, \ldots, h_\alpha^k) \xrightarrow{\text{a.s.}} \mathbb{E}\,\psi(Z^{h^1}, \ldots, Z^{h^k}),$$

  *for any vectors $h^1, \ldots, h^k \in \mathbb{R}^n$ in the same CDC, where $Z^{h^i}$ are as defined in Definition F.6.*

- *Any scalar $\theta$ in the program tends to $\mathring{\theta}$ almost surely, where $\mathring{\theta}$ is as defined in Definition F.6.*

Again, Theorem F.9 strictly generalized by the following Master Theorem for Parameter-Controlled nonlinearities.

**Theorem F.10** (Variable Dimension Parameter-Controlled NETSOR$\top^+$ Master Theorem)**.** *Fix a* NETSOR$\top$ *program (with variable dimensions). Suppose the initial matrices $\mathcal{W}$, vectors $\mathcal{V}$, and scalars $\mathcal{C}$ are sampled in the fashion of Setup F.5. Suppose the program satisfies the* rank stability assumption *(Assumption E.7). Assume $\varphi^u(-;-)$ is parameter-controlled at $\mathring{\Theta}^u$ for all vectors and scalars $u$. Then*

- *For any random vector $\Theta \in \mathbb{R}^l$ that converges almost surely to a deterministic vector $\mathring{\Theta}$, and for any $\psi(-;-) : \mathbb{R}^k \times \mathbb{R}^l \to \mathbb{R}$ parameter-controlled at $\mathring{\Theta}$, we have*

$$\frac{1}{n}\sum_{\alpha=1}^{n} \psi(g_\alpha^1, \ldots, g_\alpha^k; \Theta) \xrightarrow{\text{a.s.}} \mathbb{E}\,\psi(Z^{g^1}, \ldots, Z^{g^k}; \mathring{\Theta}).$$

  *in the limit described in Setup F.5, for any G-vars $g^1, \ldots, g^k \in \mathbb{R}^n$ in the same CDC, where $Z^{g^i}$ are as defined in Definition F.6.*

- *Any scalar $\theta$ in the program tends to $\mathring{\theta}$ almost surely, where $\mathring{\theta}$ is as defined in Definition F.6.*

# G Random Matrix Theory Background

In random matrix theory, we often are concerned with *what the spectrum of a random matrix looks like* when the matrix is large. The answer to this question is typically phrased in a convergence theorem which says *the spectrum as a histogram converges to some distribution*. We single out the notion of convergence we are concerned with in this paper.

**Definition G.1.** Let $\mu_n$ be a *random* measure (such as the spectral distribution of a random matrix) on $\mathbb{R}$ for each $n = 1, 2, \ldots$. Let $\mu$ be a *deterministic* measure on $\mathbb{R}$. We say $\mu_n$ converges almost surely to $\mu$, written $\mu_n \xrightarrow{\text{a.s.}} \mu$, if for every compactly supported continuous function $\varphi$, we have $\int_{\mathbb{R}} \varphi(x)\,\mathrm{d}\mu_n(x) \xrightarrow{\text{a.s.}} \int_{\mathbb{R}} \varphi(x)\,\mathrm{d}\mu(x)$ (where the almost sure convergence is over the randomness of $\mu_n$).

The *Moment Method* is a standard technique in probability theory to establish convergence in distribution. In the context of random matrix theory, this method goes like the following:

- Under general conditions (see Carleman's condition Fact G.3), a distribution is "pinned down" by its moments: if all moments of a distribution $\mu$ converges to the moment of a target distribution $\mu^*$, then we can in general say that $\mu \to \mu^*$ for some notion of convergence $\to$.
- For a Hermitian matrix $A$ with eigenvalues $\lambda_1, \ldots, \lambda_n$, the empirical distribution $\mu$ of its eigenvalues has moments given by $\mathbb{E}_{\lambda\sim\mu}\,\lambda^r = \frac{1}{n}\sum_{\alpha=1}^{n}(\lambda_\alpha)^r = \frac{1}{n}\operatorname{tr} A^r$.
- Thus, to prove that the spectral distribution $\mu$ of a random matrix converges to some distribution $\mu^*$, we need to show, for all $r \geq 0$, we have $\operatorname{tr} A^r \to \mathbb{E}_{\lambda\sim\mu^*}\,\lambda^r$, for some notion of convergence $\to$.

This method implies that, to check the convergence in Definition G.1, it suffices to check the convergence of moments:

**Fact G.2** (Moment Method)**.** *In the context of Definition G.1 where $\mu_n$ and $\mu$ are probability measures, $\mu_n \xrightarrow{\text{a.s.}} \mu$ if 1) $\mu$ satisfies Carleman's Condition (Fact G.3) and so does $\mu_n$ almost surely for large enough $n$, and 2) for every $k = 1, 2, \ldots$,*

$$\mathop{\mathbb{E}}_{x\sim\mu_n} x^k \xrightarrow{\text{a.s.}} \mathop{\mathbb{E}}_{x\sim\mu} x^k$$

*where the convergence $\xrightarrow{\text{a.s.}}$ is over the randomness of the measure $\mu_n$.*

For more background on the Moment Method, see [42].

**Fact G.3** (Carleman's Condition)**.** *Let $\mu$ be a measure on $\mathbb{R}$ such that all moments*

$$M_k = \int_{-\infty}^{\infty} x^n \,\mathrm{d}\mu(x), \quad k = 0, 1, 2, \ldots$$

*are finite. If*

$$\sum_{k=1}^{\infty} M_{2k}^{-\frac{1}{2k}} = \infty,$$

*then $\mu$ is the only measure on $\mathbb{R}$ with $\{M_k\}_k$ as its moments.*

# H Proving the Marchenko-Pastur Law by Tensor Programs

Marchenko-Pastur law [27] is another cornerstone random matrix result on the level of semicircle law. It says that the spectrum $\mu_{AA^\top}$ of (what's known as Wishart) matrices $AA^\top$, where $A \in \mathbb{R}^{m\times n}, A_{\alpha\beta} \sim \mathcal{N}(0, 1/n)$, converges in distribution if the shape ratio $m/n \to \rho$ for a finite but nonzero $\rho \in (0, \infty)$: (see Definition G.1 for meaning of $\xrightarrow{\text{a.s.}}$)

$$\mu_{AA^\top} \xrightarrow{\text{a.s.}} \mu_{\text{mp}}(x) \stackrel{\text{def}}{=} \mathrm{relu}(1-\rho^{-1})\delta(x) + \frac{1}{\rho 2\pi x}\sqrt{(b-x)(x-a)}\,\mathbb{I}_{[a,b]}(x)\,\mathrm{d}x \tag{26}$$

where $\delta(x)$ is the Dirac Delta, $\mathrm{relu}(x) = x\mathbb{I}(x > 0)$, $a = (1-\sqrt{\rho})^2$, and $b = (1+\sqrt{\rho})^2$. This distribution also yields the (square of) of the singular value distributions of $A$.

In this section, we use NETSOR$\top$ with variable dimensions (Appendix F) to prove this law. Like for the semicircle law, our purpose is to 1) demonstrate concrete examples of computing with NETSOR$\top$ with variable dimensions, and 2) "benchmark" and show our framework powerful enough to prove this nontrivial result.

Like in Section 3, we adopt the Moment Method (Fact G.2). This means that we need to show that the moments of $AA^\top$ converge to the corresponding moments of $\mu$:

$$\frac{1}{n}\mathrm{tr}\left(AA^\top\right)^k \to \mathop{\mathbb{E}}_{\lambda\sim\mu_{\text{mp}}} \lambda^k, \quad \text{for all } k = 1, 2, \ldots \tag{27}$$

To calculate this moment $\mathrm{tr}\left(AA^\top\right)^k$, we use the trace trick to rewrite it as

$$\frac{1}{n}\mathrm{tr}\left(AA^\top\right)^k = \frac{1}{n}\mathop{\mathbb{E}}_{v\sim\mathcal{N}(0,I)} v^\top \left(AA^\top\right)^k v.$$

We can then express the RHS in a NETSOR$\top$ program with variable dimensions: Let $\mathcal{V} = \{v\}, \mathcal{W} = \{A\}$, and introduce new vectors via `MatMul` like so

$$u^i = A^\top v^{i-1}, \quad v^i = Au^i, \quad i = 1, \ldots, k,$$

where we have set $v^0 = v$ for convenience. Then mathematically, $v^i = (AA^\top)^i v$. By Theorem 3.3,

$$\frac{1}{n}\mathrm{tr}\left(AA^\top\right)^k = \mathop{\mathbb{E}}_{v}\frac{v^\top v^k}{n} \xrightarrow{\text{a.s.}} \mathbb{E}\, Z^v Z^{v^k}.$$

Thus, we are done if can demonstrate

$$\mathbb{E}\, Z^v Z^{v^k} = \mathop{\mathbb{E}}_{\lambda\sim\mu_{\text{mp}}} \lambda^k$$

Now, all that is left is to *compute* $Z^{v^i}$.

### H.1 Explicit Calculations of the First Few Moments

To give a concrete feel for the proof, we first calculate the first few moments as examples. Before we begin, note that $\mathring{\sigma}_A^2 = 1$ and $\mathring{\sigma}_{A^\top}^2 = \rho$ (where $\mathring{\sigma}_{A^\top}^2$ is as defined in Setup F.5).

**First Moment** First we have $Z^v = \mathcal{N}(0, 1)$, $Z^{u^1} = \mathcal{N}(0, \mathring{\sigma}_{A^\top}^2) = \mathcal{N}(0, \rho)$, independent of each other. Then

$$Z^{v^1} = \hat{Z}^{v^1} + \dot{Z}^{v^1},$$

where $\dot{Z}^{v^1} = \mathring{\sigma}_A^2 Z^v \mathbb{E} \frac{\partial Z^{u^1}}{\partial \hat{Z}^{u^1}} = Z^v$ and $\hat{Z}^{v^1} = \mathcal{N}(0, \rho)$ independent of everything else. Thus, the first moment of $AA^\top$ has limit

$$\frac{1}{n} \operatorname{tr} AA^\top \xrightarrow{\text{a.s.}} \mathbb{E} Z^v Z^{v^1} = \mathbb{E} Z^v \dot{Z}^{v^1} = \mathbb{E}(Z^v)^2 = 1 = \mathop{\mathbb{E}}_{\lambda \sim \mu_{\text{mp}}} \lambda$$

**Second Moment** Next, we have $Z^{u^2} = Z^{A^\top v^1} = \hat{Z}^{u^2} + \dot{Z}^{u^2}$ where

$$\dot{Z}^{u^2} = \mathring{\sigma}_{A^\top}^2 \left( Z^{u^1} \mathbb{E} \frac{\partial Z^{v^1}}{\partial \hat{Z}^{v^1}} \right) = \rho Z^{u^1},$$

and $\hat{Z}^{u^2}$ is zero-mean and jointly Gaussian with $\hat{Z}^{u^1}$, with

$$\begin{aligned}
\operatorname{Var}\left(\hat{Z}^{u^2}\right) &= \mathring{\sigma}_{A^\top}^2 \mathbb{E}\left(Z^{v^1}\right)^2 = \rho\left(\mathbb{E}\left(\hat{Z}^{v^1}\right)^2 + \mathbb{E}\left(\dot{Z}^{v^1}\right)^2\right) = \rho(\rho+1) = \rho^2 + \rho \\
\operatorname{Cov}\left(\hat{Z}^{u^2}, \hat{Z}^{u^1}\right) &= \mathring{\sigma}_{A^\top}^2 \mathbb{E} Z^{v^1} Z^v = \rho \mathbb{E} \dot{Z}^{v^1} Z^v = \rho.
\end{aligned}$$

Then $Z^{v^2} = Z^{Au^2} = \hat{Z}^{v^2} + \dot{Z}^{v^2}$, where

$$\dot{Z}^{v^2} = \mathring{\sigma}_A^2 \left( Z^{v^1} \mathbb{E} \frac{\partial Z^{u^2}}{\partial \hat{Z}^{u^2}} + Z^v \mathbb{E} \frac{\partial Z^{u^2}}{\partial \hat{Z}^{u^1}} \right) = Z^{v^1} + \rho Z^v.$$

Thus, the second moment of $AA^\top$ has limit

$$\frac{1}{n} \operatorname{tr}\left(AA^\top\right)^2 \xrightarrow{\text{a.s.}} \mathbb{E} Z^v Z^{v^2} = \mathbb{E} Z^v \left(Z^{v^1} + \rho Z^v\right) = 1 + \rho = \mathop{\mathbb{E}}_{\lambda \sim \mu_{\text{mp}}} \lambda^2.$$

### H.2 Proof for General Moments

In general, we have

- $\{\hat{Z}^{v^i}\}_i$ is jointly Gaussian with covariance
$$\operatorname{Cov}\left(\hat{Z}^{v^i}, \hat{Z}^{v^j}\right) = \mathbb{E} Z^{u^i} Z^{u^j},$$
- $\{\hat{Z}^{u^i}\}_i$ is jointly Gaussian with covariance
$$\operatorname{Cov}\left(\hat{Z}^{u^i}, \hat{Z}^{u^j}\right) = \rho \mathbb{E} Z^{v^{i-1}} Z^{v^{j-1}},$$
- and
$$\dot{Z}^{v^i} = \sum_{j=0}^{i-1} Z^{v^j} \mathbb{E} \frac{\partial Z^{u^i}}{\partial \hat{Z}^{u^{j+1}}}, \quad \dot{Z}^{u^i} = \rho \sum_{j=1}^{i-1} Z^{u^j} \mathbb{E} \frac{\partial Z^{v^{i-1}}}{\partial \hat{Z}^{v^j}}. \tag{28}$$

Expanding all $Z^\bullet$ into $\hat{Z}^\bullet + \dot{Z}^\bullet$ recursively, it is easy to see that there are coefficients $\{a_j^i\}_{i,j}$, $\{b_j^i\}_{i,j}$ such that

$$Z^{v^i} = \sum_{j=0}^{i} a_j^i \hat{Z}^{v^j}, \quad Z^{u^i} = \sum_{j=1}^{i} b_j^i \hat{Z}^{u^j}. \tag{29}$$

Since $\hat{Z}^v = \hat{Z}^{v^0}$ is independent from all other $\hat{Z}^\bullet$, we thus seek to prove

$$a_0^k = \mathbb{E}\, Z^{v^k} Z^v = \mathop{\mathbb{E}}_{\lambda \sim \mu_{\mathrm{mp}}} \lambda^k.$$

Given Eq. (29), we see

$$a_j^i = \mathbb{E}\, \frac{\partial Z^{v^i}}{\partial \hat{Z}^{v^j}}, \quad b_j^i = \mathbb{E}\, \frac{\partial Z^{u^i}}{\partial \hat{Z}^{u^j}}.$$

Plugging Eq. (29) into Eq. (28), we see

$$\begin{aligned}
\dot{Z}^{v^i} &= \sum_{j=0}^{i-1} Z^{v^j}\, \mathbb{E}\, \frac{\partial Z^{u^i}}{\partial \hat{Z}^{u^{j+1}}} = \sum_{j=0}^{i-1} Z^{v^j} b_{j+1}^i \\
&= \sum_{j=0}^{i-1} \left( \sum_{l=0}^{j} a_l^j \hat{Z}^{v^l} \right) b_{j+1}^i = \sum_{l=0}^{i-1} \hat{Z}^{v^l} \sum_{j=l}^{i-1} a_l^j b_{j+1}^i.
\end{aligned}$$

Then

$$\sum_{l=0}^{i} a_l^i \hat{Z}^{v^l} = Z^{v^i} = \hat{Z}^{v^i} + \dot{Z}^{v^i} = \hat{Z}^{v^i} + \sum_{l=0}^{i-1} \hat{Z}^{v^l} \sum_{j=l}^{i-1} a_l^j b_{j+1}^i.$$

Matching coefficients, we obtain the recurrence relation

$$a_i^i = 1, \quad \text{and for all } l = 0, \ldots, i-1, \quad a_l^i = \sum_{j=l}^{i-1} a_l^j b_{j+1}^i. \tag{30}$$

Similarly

$$\begin{aligned}
\dot{Z}^{u^i} &= \rho \sum_{j=1}^{i-1} Z^{u^j}\, \mathbb{E}\, \frac{\partial Z^{v^{i-1}}}{\partial \hat{Z}^{v^j}} = \rho \sum_{j=1}^{i-1} Z^{u^j} a_j^{i-1} \\
&= \rho \sum_{j=1}^{i-1} \left( \sum_{l=1}^{j} b_l^j \hat{Z}^{u^l} \right) a_j^{i-1} = \rho \sum_{l=1}^{i-1} \hat{Z}^{u^l} \sum_{j=l}^{i-1} b_l^j a_j^{i-1}.
\end{aligned}$$

Then

$$\sum_{l=1}^{i} b_l^i \hat{Z}^{u^l} = Z^{u^i} = \hat{Z}^{u^i} + \dot{Z}^{u^i} = \hat{Z}^{u^i} + \rho \sum_{l=1}^{i-1} \hat{Z}^{u^l} \sum_{j=l}^{i-1} b_l^j a_j^{i-1}.$$

Matching coefficients, we get

$$b_i^i = 1, \quad \text{and for all } l = 1, \ldots, i-1, \quad b_l^i = \rho \sum_{j=l}^{i-1} b_l^j a_j^{i-1}. \tag{31}$$

Let $M_r \stackrel{\text{def}}{=} \mathbb{E}_{\lambda \sim \mu_{\mathrm{mp}}} \lambda^r$. Then we claim that the solution to the recurrence Eq. (30) and Eq. (31) is given by

$$a_j^i = M_{i-j}, \quad b_i^i = 1, \quad \text{and} \quad b_j^i = \rho M_{i-j} \quad \text{for all } i - j \geq 1.$$

Plugging into Eq. (30) and Eq. (31), we see it remains to show the following Catalan-like identity

$$M_s = \rho \sum_{r=1}^{s-2} M_r M_{s-1-r} + (1 + \rho) M_{s-1}. \tag{32}$$

This can be done by a change of variables to express $\mathbb{E}_{\lambda \sim \mu_{\mathrm{mp}}} \lambda^r$ as an expectation over the semicircle law $\mu_{\mathrm{sc}}$:

$$M_r = \mathop{\mathbb{E}}_{\lambda \sim \mu_{\mathrm{mp}}} \lambda^r = \mathop{\mathbb{E}}_{\lambda \sim \mu_{\mathrm{sc}}} (1 + \rho + \lambda \sqrt{\rho})^{r-1}.$$

Expanding the power and using the Catalan identity Eq. (5) for the moments of the semicircle law, we have

$$M_r = \sum_{k=0}^{\lfloor (r-1)/2 \rfloor} \alpha^k (1 + \alpha)^{r-1-2k} \binom{r-1}{2k} C_k \tag{33}$$

where $C_k$ is the $k$th Catalan number. Then one can verify Eq. (32) by expanding $M_r$ into Catalan numbers using Eq. (33), and repeatedly applying the Catalan identity $C_k = \sum_{i=0}^{k-1} C_i C_{k-1-i}$. This finishes the proof of the Marchenko-Pastur law.

# I Subprogram and Subprogram Independence

To prove Theorem 4.2, we need to introduce a notion of *subprograms* and then the *independence* of subprograms. Our key lemma in this section is Lemma I.5, which shows that, if a subprogram is sufficiently uncorrelated from the previous subprogram, then it is also independent from it.

**Definition I.1.** A program can be written down formally as a sequence of lines, each assigning a new vector generated from `Nonlin` or `MatMul` to a new variable, like the examples above. A *subprogram* of a program is a span of such lines. Given a program $\pi$ and subprograms $\pi_1, \ldots, \pi_k$, we write $\pi = \pi_1|\cdots|\pi_k$ if $\pi$ consists of $\pi_1$ followed by $\pi_2$ followed by $\pi_3$, and so on, ending with $\pi_k$. Each vector in $\pi$ is introduced in a unique subprogram $\pi_i$, for which we write $v \in \pi_i$. By convention, all initial vectors in $\mathcal{V}$ are included in the first subprogram $\pi_1$.

For example, the simple NETSOR$\top$ program represented by Eq. (2) can be split into two subprograms, one which introduces $x^l, h^l$ for $l = 1, \ldots, R$, and another with $l = R+1, \ldots, L$. As another example, Eq. (21) and Eq. (22) are two subprograms that comprise the program expressing the forward and backward propagations of the network. In general, forward propagation and backward propagation form natural subprograms, where the latter depends on (the preactivations computed in) the former.

Next, we introduce the notion of subprogram independence, which generalizes the main property of Simple GIA Check (Condition 1). Before that, we first recall the notation for distributional equality.

**Notations** Given two random variables $X, Z$, and a $\sigma$-algebra $\mathcal{A}$, the notation $X \overset{\mathrm{d}}{=}_{\mathcal{A}} Z$ means that for any integrable function $\phi$ and for any random varible $Z$ measurable on $\mathcal{A}$, $\mathbb{E}\,\phi(X)Z = \mathbb{E}\,\phi(Z)Z$. We say that $X$ is distributed as (or is equal in distribution to) $Z$ conditional on $\mathcal{A}$. In case $\mathcal{A}$ is the trivial $\sigma$-algebra, we just write $X \overset{\mathrm{d}}{=} Z$. If $X$ and $Z$ agrees almost surely, then we write $X \overset{\mathrm{a.s.}}{=} Z$.

**Definition I.2.** Consider a program divided into two subprograms: $\pi = \pi_1|\pi_2$. For each $W \in \mathcal{W}$, we create an iid copy $\bar{W}$. Let $\bar{\pi}_2$ be the same as $\pi_2$ except that each matrix $W$ used in $\pi_2$ is replaced with $\bar{W}$ in $\bar{\pi}_2$. Let $x^1, \ldots, x^k$ denote the vectors of $\pi_1$, $y^1, \ldots, y^p$ denote the vectors of $\pi_2$, and $\bar{y}^1, \ldots, \bar{y}^p$ denote the vectors of $\bar{\pi}_2$. We say $\pi_2$ *is independent from* $\pi_1$ if we have the distributional equality

$$(Z^{x^1}, \ldots, Z^{x^k}, Z^{y^1}, \ldots, Z^{y^p}) \overset{\mathrm{d}}{=} (Z^{x^1}, \ldots, Z^{x^k}, Z^{\bar{y}^1}, \ldots, Z^{\bar{y}^p})$$

or equivalently, the conditional distributional equality

$$(Z^{y^1}, \ldots, Z^{y^p}) \overset{\mathrm{d}}{=}_{Z^{x^1}, \ldots, Z^{x^k}} (Z^{\bar{y}^1}, \ldots, Z^{\bar{y}^p}).$$

For example, if $\pi_1$ expresses the forward computation of a network satisfying Simple GIA Check (Condition 1), and $\pi_2$ expresses the backward computation, then we can assume $W^\top$ used in $\pi_2$ (backpropagation) is independent from $W$ used in $\pi_1$ (forward propagation). This then implies $\pi_2$ is independent from $\pi_1$. On the other hand, the subprogram given by Eq. (22) is *not* independent from the subprogram given by Eq. (21), as we calculated.

Note Definition I.2 is purely about the structure of the subprograms (which is the only thing the random variables $Z^\bullet$ depend on), not the values of the vectors or matrices in the program.

We will talk about polynomials of matrices below, e.g. $WW^\top W^2(W^\top)^2$ if $W$ is a square matrix.

**Definition I.3** (Noncommutative Polynomial)**.** Given a matrix $W \in \mathbb{R}^{m\times n}$, a *(noncommutative) monomial* in $\{W, W^\top\}$ is a product $aX_1X_2\cdots X_k$ for some $a \in \mathbb{R}$, $k \geq 0$, where each $X_i \in \{W, W^\top\}$ (with appropriate matching shapes). Its degree is $k$. A *(noncommutative) polynomial* in $\{W, W^\top\}$ is a linear combination of such monomials. Its degree is the maximum degree of all of its nonzero monomials.

*Remark* I.4. If $W$ and $v$ are a matrix and a vector of a program, and $A$ is a polynomial in $\{W, W^\top\}$, then one can naturally express the product $Av$ in NETSOR$\top$. For example, if $A = W(W^\top)^2W$, then $Av = u$ where

$$v^1 = Wv, \quad v^2 = W^\top v^1, \quad v^3 = W^\top v^2, \quad u = Wv^3.$$

Below, when we say a program *calculates* $Av$, we mean a program of this form.

The following is the key workhorse of this section. It says that, in a program $\pi_1|\pi_2$, if $\pi_2$ 1) consists of only `MatMul` and 2) only depends on $\pi_1$ through a single vector $v$ *uncorrelated* with most vectors of $\pi_1$, then $\pi_2$ is independent from $\pi_1$. As we will comment shortly, this gives another proof of the Gradient Independence Assumption when Simple GIA Check (Condition 1) is satisfied.

**Lemma I.5** (Independence from Uncorrelation)**.** *Consider a* NETSOR$\top$ *program* $\pi = \pi_1|\pi_2$*, and* $\pi_2$ *calculates* $Av$ *(in the sense of Remark I.4) where* $v$ *is a vector of* $\pi_1$ *and* $A$ *is a polynomial in* $\{W, W^\top\}$ *for some* $W \in \mathcal{W}$*. If* $\mathbb{E} Z^v Z^x = \mathbb{E} Z^v \hat{Z}^y = 0$ *for every* $x, y \in \pi_1$ *with* $y = Wx$ *or* $y = W^\top x$*, then* $\pi_2$ *is independent from* $\pi_1$*.*

Note if $Wv$ or $W^\top v$ appears in $\pi_1$, then the premise implies $Z^v = 0$ and the theorem becomes vacuously true. So to get nontrivial implications, $Wv$ and $W^\top v$ should not appear in $\pi_1$. Given Lemma I.5, we can obtain a more explicit form of $Z^{Av}$ by some straightforward calculation.

**Lemma I.6.** *In the setting of Lemma I.5, we have*

$$Z^{Av} = \tau Z^v + S$$

*where* $S$ *is zero-mean and independent from* $\{Z^u : u \in \pi_1\}$*, and* $\tau = \lim_{n\to\infty} \frac{1}{n} \operatorname{tr} A$*.*

*Proof.* By Lemma I.5, we can assume that $W$ (where $A$ is a polynomial in $\{W, W^\top\}$) is independent from everything in $\pi_1$. Then a simple induction on the degree of $A$ in $W$ shows that

$$Z^{Av} = \tau Z^v + S \tag{34}$$

where $S$ is zero-mean and independent from $\{Z^u : u \in \pi_1\}$.

However, at this point, we don't know what $\tau$ would be. To understand $\tau$, we proceed as follows. Let $\bar{v}$ be sampled iid according to $\bar{v}_\alpha \sim Z^v$ (which can be constructed with a NETSOR$\top$ program[40]), so that $Z^{\bar{v}} \overset{\mathrm{d}}{=} Z^v$. Then we can easily see from the definition of $Z$ that $Z^{A\bar{v}} = \tau Z^{\bar{v}} + \bar{S}$ where $\tau$ here is the same as $\tau$ in Eq. (34) and $(Z^{\bar{v}}, \bar{S}) \overset{\mathrm{d}}{=} (Z^v, S)$. Then by the Master Theorem,

$$\tau\, \mathbb{E}(Z^{\bar{v}})^2 = \mathbb{E}\, Z^{A\bar{v}} Z^{\bar{v}} = \lim_{n\to\infty} \frac{1}{n} \bar{v}^\top A \bar{v} = \lim_{n\to\infty} \frac{1}{n} \operatorname{tr} A\, \mathbb{E}(Z^{\bar{v}})^2$$

we get $\tau = \lim_{n\to\infty} \frac{1}{n} \operatorname{tr} A$ as desired. $\square$

**Extension to NETSOR$\top^+$ and Variable Dimensions** Both Lemmas I.5 and I.6 hold as stated for NETSOR$\top^+$ programs (see Appendix E) and programs with variable dimensions (see Appendix F). Note that, in programs with variable dimensions, the $\tau$ in Lemma I.6 vanishes if $v$ and $Av$ are in different CDCs.

**Another Look at How Simple GIA Check Implies GIA** The lemmas above give us another proof that the gradient independence assumption leads to correct calculations given Simple GIA Check (Theorem D.1). To demonstrate Lemmas I.5 and I.6 in action, before proving FIP, we give a proof of Theorem D.1 for MLP.

*Proof sketch for an MLP.* Consider the MLP $f(\xi) = W^{L+1} x^L(\xi)$ with input $\xi \in \mathbb{R}^d$ and output dimension 1, where we recursively define, for $l = 1, \ldots, L$,

$$\begin{aligned} g^l(\xi) &= W^l x^{l-1}(\xi) & \qquad dx^{l-1}(\xi) &= W^{l\top} dg^l(\xi) \\ x^l(\xi) &= \phi(g^l(\xi)) & \qquad dg^l(\xi) &= \phi'(g^l(\xi)) \odot dx^l(\xi). \end{aligned}$$

Here, the dimensionalities are $W^1 \in \mathbb{R}^{n\times d}$; $W^2, \ldots, W^L \in \mathbb{R}^{n\times n}$; $W^{L+1} \in \mathbb{R}^{1\times n}$, and for all $l \in [L]$, $g^l, x^l, dx^l, dg^l \in \mathbb{R}^n$.

Let $\pi_F$ be the program that computes $g^l, x^l, l = 1, \ldots, L$, on an input $\xi$. Below, we abbreviate $g^l = g^l(\xi)$ and so on. In this program, $\mathcal{V} = \{g^1, dx^L = \sqrt{n} W^{L+1}\}$, [41]and $\mathcal{W} = \{W^2, \ldots, W^L\}$.[42]

[40] The vector $v$ can always be expressed as $\phi(g^1, \ldots, g^k)$ for some G-vars $g^1, \ldots, g^k$ and some function $\phi : \mathbb{R}^k \to \mathbb{R}$. Then $\bar{v}$ can be constructed as $\phi(\bar{g}^1, \ldots, \bar{g}^k)$ where $\{\bar{g}^1_\alpha, \ldots, \bar{g}^k_\alpha\} \sim \mathcal{N}(\mu, \Sigma)$ with $\mu_i = \mathbb{E} Z^{g^i}$ and $\Sigma_{ij} = \mathbb{E} Z^{g^i} Z^{g^k}$.

[41] Here $dx^l$ should be interpreted as $\sqrt{n}\partial f/\partial x^l$, and likewise for $dg^l$.

[42] For concreteness (which will not matter much below): we sample $W^l_{\alpha\beta} \sim \mathcal{N}(0, 1/n)$ for $l = 2, \ldots, L+1$, and suppose $g^1_\alpha \overset{\mathrm{d}}{=} \mathcal{N}(0, 1)$, induced by appropriate sampling of $W^1$.

Now let's see how to apply Lemmas I.5 and I.6 to show Theorem D.1. The first step of the backward pass is $dg^L = \phi'(g^L) \odot dx^L$. Let $\pi_1$ denote $\pi_F$ plus this line. Let $\pi_2$ be the next line $dx^{L-1} = W^{L\top} dg^L$. Then $\pi_1, \pi_2$ along with $v = dg^L$ satisfies the condition of Lemma I.5. Indeed, $Z^{dg^L}$ is uncorrelated from $Z^{g^L}$ and $Z^{x^{L-1}} = \hat{Z}^{x^{L-1}}$ because $Z^{dg^L}$ is linear in $Z^{dx^L}$, which is by definition sampled independently from all $W^1, \ldots, W^L$. Thus, by Lemma I.5, we may *rigorously* treat $W^{L\top}$ as independent from $W^L$ for the purpose of computing any quantity of the form of the Theorem 2.10 for the program $\pi_1|\pi_2$. In particular, by Lemma I.6, $Z^{dx^{L-1}} = \tau Z^{dg^L} + S$, where $S$ is zero-mean and independent from $\{Z^u : u \in \pi_1\}$, and $\tau = \lim_{n\to\infty} \frac{1}{n} \operatorname{tr} W^{L\top} = 0$, i.e. $Z^{dx^{L-1}} = S$.

Then we can repeat this reasoning inductively: $Z^{dg^{L-1}} = \phi'(Z^{g^{L-1}}) \cdot Z^{dx^{L-1}} = \phi'(Z^{g^{L-1}}) \cdot S$ is uncorrelated from $Z^{g^{L-1}}$ and $Z^{x^{L-2}} = \hat{Z}^{x^{L-2}}$ because $Z^{dg^{L-1}}$ is linear in $S$. This lets us apply Lemmas I.5 and I.6 to $\pi_2$ being the line $dx^{L-2} = W^{L-1\top} dg^{L-1}$, $\pi_1$ being all previous lines, and $v$ being $dg^{L-1}$. So on and so forth. ☐

### I.1 Proving Lemma I.5

Here we prove Lemma I.5. It will use the following trivial but useful facts.

**Proposition I.7.** *For any $x \in \pi$, $Z^x$ is measurable against the $\sigma$-algebra generated by $\{\hat{Z}^y :$ G-var $y \in \pi\}$. .*

**Lemma I.8.** *Fix a matrix $W$. If $\mathbb{E}\, Z^v \hat{Z}^{Wh} = 0$ for all G-var $Wh$ introduced before $v$, then $\mathbb{E}\, \partial Z^v / \partial \hat{Z}^{Wh} = 0$ for all such $Wh$ as well.*

Proposition I.7 follows from a simple inductive argument, and Lemma I.8 follows from Eq. (4).

*Proof of Lemma I.5.* It suffices to prove the case when we assume $A = \prod_{i=1}^p W^{t_i}, t_i \in \{1, \top\}$, can be expressed as a monomial of degree $p$ in $\{W, W^\top\}$. Let $\bar{W}$ be an iid copy of $W$ as in Definition I.2 and likewise let $\bar{\pi}_2, y^1, \ldots, y^p, \bar{y}^1, \ldots, \bar{y}^p$ be as in Definition I.2. By our assumption, $\pi_2$ computes $y^1 = W^{t_1} v$ and $y^i = W^{t_i} y^{i-1}$ for $i = 2, \ldots, p$, and likewise $\bar{\pi}_2$ computes $\bar{y}^1 = \bar{W}^{t_1} v$ and $\bar{y}^i = \bar{W}^{t_i} \bar{y}^{i-1}$ for $i = 2, \ldots, p$.

Define $\mathcal{X}$ to be the $\sigma$-algebra generated by $\{Z^x : x \in \pi_1\}$. Note that by Proposition I.7, $\mathcal{X}$ is also the $\sigma$-algebra generated by $\{\hat{Z}^x :$ G-var $x \in \pi_1\}$. Then we need to show

$$(Z^{y^1}, \ldots, Z^{y^p}) \overset{\mathrm{d}}{=}_{\mathcal{X}} (Z^{\bar{y}^1}, \ldots, Z^{\bar{y}^p}).$$

We proceed by induction on $p$. The base case of $p = 0$ is vacuously true.

Suppose the induction hypothesis holds for $p = q - 1$

$$(Z^{y^1}, \ldots, Z^{y^{q-1}}) \overset{\mathrm{d}}{=}_{\mathcal{X}} (Z^{\bar{y}^1}, \ldots, Z^{\bar{y}^{q-1}}).$$

and we shall prove it for $p = q$. A simple inductive argument (as in our proof of the semicircle law; see Section 3.3) shows that for each $i \in [q-1]$, we have

$$Z^{\bar{y}^i} = \tau_i Z^v + \bar{S}_i \tag{35}$$

for some $\tau_i \in \mathbb{R}$ and some zero-mean $\bar{S}_i$ independent from $\mathcal{X}$ (and thus also independent from $Z^v$).

Since we have assumed $A$ to be a monomial in $\{W, W^\top\}$, we can suppose WLOG that $t_q = 1$ so that $y^q = W y^{q-1}$. By definition, $Z^{y^q} = \hat{Z}^{y^q} + \dot{Z}^{y^q}$ and $Z^{\bar{y}^q} = \hat{Z}^{\bar{y}^q} + \dot{Z}^{\bar{y}^q}$. We shall show

$$(Z^{y^1}, \ldots, Z^{y^{q-1}}, \hat{Z}^{y^q}, \dot{Z}^{y^q}) \overset{\mathrm{d}}{=}_{\mathcal{X}} (Z^{\bar{y}^1}, \ldots, Z^{\bar{y}^{q-1}}, \hat{Z}^{\bar{y}^q}, \dot{Z}^{\bar{y}^q})$$

which would imply the IH for $p = q$.

**Case $\hat{Z}^{y^q}$:** By definition, $\hat{Z}^{y^q}$ (resp. $\hat{Z}^{\bar{y}^q}$) is zero-mean, jointly Gaussian with $\{\hat{Z}^{Wx} : x, Wx \in \pi\}$ (resp. $\{\hat{Z}^{\bar{W}x} : x, \bar{W}x \in \bar{\pi}_2\}$), and independent from $\{\hat{Z}^{Qx} : x, Qx \in \pi\}$ for any $Q \neq W$ (resp. for any $Q \neq \bar{W}$). By Proposition I.7, these facts fully determine the distributions of $Z^{y^1}, \ldots, Z^{y^{q-1}}, \hat{Z}^{y^q}$ and of $Z^{\bar{y}^1}, \ldots, Z^{\bar{y}^{q-1}}, \hat{Z}^{\bar{y}^q}$ conditioned on $\mathcal{X}$. *We thus need to show that A)* $\mathrm{Cov}(\hat{Z}^{y^q}, \hat{Z}^{y^i}) = \mathrm{Cov}(\hat{Z}^{\bar{y}^q}, \hat{Z}^{\bar{y}^i})$ *for all $i = 1, \ldots, q-1$, and B)* $\mathrm{Cov}(\hat{Z}^{y^q}, \hat{Z}^g) = \mathrm{Cov}(\hat{Z}^{\bar{y}^q}, \hat{Z}^g)$ *for all G-var $g \in \pi_1$.*

**A)** For any $i = 2, \ldots, q-1$, we either have $y^i = Wy^{i-1}$ or $y^i = W^\top y^{i-1}$. 1) In the latter case, $\hat{Z}^{y^i}$ is independent from $\hat{Z}^{y^q}$, and likewise $\hat{Z}^{\bar{y}^i}$ is independent from $\hat{Z}^{\bar{y}^q}$, so trivially $\mathrm{Cov}(\hat{Z}^{y^q}, \hat{Z}^{y^i}) = \mathrm{Cov}(\hat{Z}^{\bar{y}^q}, \hat{Z}^{\bar{y}^i}) = 0$. 2) In the former case,

$$\begin{aligned}
\mathrm{Cov}(\hat{Z}^{y^q}, \hat{Z}^{y^i}) &= \mathrm{Cov}(\hat{Z}^{Wy^{q-1}}, \hat{Z}^{Wy^{i-1}}) \\
&= \sigma_W^2 \mathbb{E}\, Z^{y^{q-1}} Z^{y^{i-1}} = \sigma_W^2 \mathbb{E}\, Z^{\bar{y}^{q-1}} Z^{\bar{y}^{i-1}} \quad \text{by IH} \\
&= \mathrm{Cov}(\hat{Z}^{\bar{y}^q}, \hat{Z}^{\bar{y}^i})
\end{aligned}$$

as desired. For $i = 1$, we either have $y^1 = Wv$ or $y^1 = W^\top v$, and a similar logic shows $\mathrm{Cov}(\hat{Z}^{y^q}, \hat{Z}^{y^1}) = \mathrm{Cov}(\hat{Z}^{\bar{y}^q}, \hat{Z}^{\bar{y}^1})$ as well.

**B)** If $g \in \pi_1$ is a G-var with $g = Qx$ for some $Q \neq W$ and $x \in \pi_1$, then by definition, $\hat{Z}^g$ is independent from both $\hat{Z}^{y^q}$ and $\hat{Z}^{\bar{y}^q}$. Now suppose instead $g = Wx$ for some $x \in \pi_1$. Then

$$\begin{aligned}
\mathrm{Cov}(\hat{Z}^{y^q}, \hat{Z}^g) &= \sigma_W^2 \mathbb{E}\, Z^{y^{q-1}} Z^x = \sigma_W^2 \mathbb{E}\, Z^{\bar{y}^{q-1}} Z^x \quad \text{by IH} \\
&= \sigma_W^2 (\tau_{q-1} \mathbb{E}\, Z^v Z^x + \mathbb{E}\, \bar{S}_{q-1} Z^x) = 0 \quad \text{by Eq. (35)} \\
&= \mathrm{Cov}(\hat{Z}^{\bar{W}\bar{y}^{q-1}}, \hat{Z}^{Wx}) = \mathrm{Cov}(\hat{Z}^{\bar{y}^q}, \hat{Z}^g).
\end{aligned}$$

Here $\mathbb{E}\, \bar{S}_{q-1} Z^x = 0$ because $\bar{S}_{q-1}$ is zero-mean and independent from $Z^x$ as $x \in \pi_1$, and $\mathbb{E}\, Z^v Z^x = 0$ by the premise of this theorem.

**Case $\dot{Z}^{y^q}$:** By definition, $\dot{Z}^{y^q}$ is a linear combination $\dot{Z}^{y^q} = \sigma_W^2 \sum_{(y,x)\in\mathcal{P}} Z^x\, \mathbb{E} \frac{\partial Z^{y^{q-1}}}{\partial \hat{Z}^y}$, where $\mathcal{P}$ is the set of all $y = W^\top x$ introduced before $y^q$ in $\pi$. Likewise, $\dot{Z}^{\bar{y}^q}$ is a linear combination $\dot{Z}^{\bar{y}^q} = \sigma_W^2 \sum_{i\in\mathcal{I}} Z^{\bar{y}^{i-1}}\, \mathbb{E} \frac{\partial Z^{\bar{y}^{q-1}}}{\partial \hat{Z}^{\bar{y}^i}}$ where $\mathcal{I}$ is the set of all $i \in [q-1]$ such that $\bar{y}^i = \bar{W}^\top \bar{y}^{i-1}$, and for convenience here we have set $y^0 = \bar{y}^0 = v$. Clearly, by IH,

$$(Z^{\bar{y}^1}, \ldots, Z^{\bar{y}^{q-1}}, \hat{Z}^{\bar{y}^q}, \dot{Z}^{\bar{y}^q}) \overset{\mathrm{d}}{=}_{\mathcal{X}} \left( Z^{y^1}, \ldots, Z^{y^{q-1}}, \hat{Z}^{y^q}, \sigma_W^2 \sum_{i\in\mathcal{I}} Z^{y^{i-1}}\, \mathbb{E} \frac{\partial Z^{y^{q-1}}}{\partial \hat{Z}^{y^i}} \right).$$

Note that $\{(y^i, y^{i-1}) : i \in \mathcal{I}\} \subseteq \mathcal{P}$. So it suffices to show that $\mathbb{E} \frac{\partial Z^{y^{q-1}}}{\partial \hat{Z}^y} = 0$ for all $(y, x) \in \mathcal{P} \setminus \{(y^i, y^{i-1}) : i \in \mathcal{I}\}$. Fix one such $y$, which must have been introduced in $\pi_1$.

Now by Eq. (35), $\mathbb{E} \frac{\partial Z^{y^{q-1}}}{\partial \hat{Z}^y} = \mathbb{E} \frac{\partial Z^{\bar{y}^{q-1}}}{\partial \hat{Z}^y} = \mathbb{E}\, \tau_{q-1} \frac{\partial Z^v}{\partial \hat{Z}^y} + \frac{\partial \bar{S}_{q-1}}{\partial \hat{Z}^y}$. Since $\bar{S}_{q-1}$ is zero-mean and independent from $\mathcal{X} \ni \hat{Z}^y$, we have $\mathbb{E} \frac{\partial \bar{S}_{q-1}}{\partial \hat{Z}^y} = 0$. At the same time, by Lemma I.8, $\mathbb{E} \frac{\partial Z^v}{\partial \hat{Z}^y} = 0$ because $\mathbb{E}\, Z^v \hat{Z}^{W^\top h} = 0$ for all G-var $W^\top h \in \pi_1$ by the premise of this theorem. Combining these results together, we get $\mathbb{E} \frac{\partial Z^{y^{q-1}}}{\partial \hat{Z}^y} = 0$ as desired. $\square$

# J Proving Free Independence Principle

*Proof of Theorem 4.2.* For each $i = 1, \ldots, t$, let $A^i$ be a polynomial in $\{W, W^\top\}$ for some $W \in \mathcal{W}$ or in $\{\mathrm{Diag}(x) : x \in \pi\}$, such that consecutive $A^i$ and $A^{i+1}$ are always polynomials in different collections. (Here superscripts always denote indices). Then we need to show

$$\frac{1}{n} \mathrm{tr}(A^t - I \frac{1}{n} \mathrm{tr}\, A^t) \cdots (A^1 - I \frac{1}{n} \mathrm{tr}\, A^1) \xrightarrow{\text{a.s.}} 0.$$

We apply the trace trick to re-express this as

$$\frac{1}{n} \mathbb{E}\, v^\top (A^t - I \frac{1}{n} \mathbb{E}\, u^{t\top} A^t u^t) \cdots (A^1 - I \frac{1}{n} \mathbb{E}\, u^{1\top} A^1 u^1) v \xrightarrow{\text{a.s.}} 0 \tag{36}$$

where expectation is taken over $v, u^1, \ldots, u^t \sim \mathcal{N}(0, I)$. We will express this as a NETSOR$\top^+$ program and apply Theorem A.7 to show this convergence. This NETSOR$\top^+$ program takes the form of $\pi|\pi'$ where $\pi$ is the program in the theorem statement defining $W \in \mathcal{W}$ and $x \in \pi$, and $\pi'$ computes $\frac{1}{n} v^\top (A^t - \frac{1}{n} u^{t\top} A^t u^t) \cdots (A^1 - \frac{1}{n} u^{1\top} A^1 u^1) v$, as we describe in more detail below[43]. .

[43] For concreteness (which won't matter in the proof below): The initial set of vectors $\mathcal{V}_{\pi|\pi'}$ is then the initial vectors of $\pi$ along with $\{v, u^1, \ldots, u^t\}$, $\mathcal{V}_{\pi|\pi'} = \mathcal{V}_\pi \cup \{v, u^1, \ldots, u^t\}$. Likewise, the initial matrices $\mathcal{W}_{\pi|\pi'}$ are the same as those of $\pi$, $\mathcal{W}_\pi$.

Inside $\pi'$ , we let $v^0 = v$ and compute $v^i$ as follows in subprograms of $\pi'$:

$$v^i = (A^i - \frac{1}{n} u^{i\top} A^i u^i) v^{i-1}, \quad \forall i \in [t].$$

Then the quantity in Eq. (36) is just $\frac{1}{n} v^\top v^t$.

**Example** If $A^i = (W^\top)^2 W$ for some $W \in \mathcal{W}$ then we can unpack $v^i = (A^i - \frac{1}{n} u^{i\top} A^i u^i) v^{i-1}$ into the subprogram

$$\begin{aligned} &g^1 = W v^{i-1} \qquad && g^2 = W^\top g^2 \qquad && g^3 = W^\top g^3 = A^i v^{i-1} \qquad && h^1 = W u^i \\ &h^2 = W^\top h^1 && h^3 = W^\top h^2 && c = \frac{1}{n} u^{i\top} h^3 && v^i = g^3 - c \cdot v^{i-1}. \end{aligned}$$

Here $c$ is computed using `Moment` and $v^i$ is computed using `Nonlin`$^+$, and everything else uses `MatMul`. Similarly, if $A^i = \mathrm{Diag}(x)$ for some $x \in \pi$ with bounded coordinates, then we can unpack $v^i = (A^i - \frac{1}{n} u^{i\top} A^i u^i) v^{i-1}$ into

$$g = x \odot v^{i-1} \qquad h = x \odot u^i \qquad c = \frac{1}{n} u^{i\top} h \qquad v^i = g - c \cdot v.$$

Here $c$ is computed using `Moment` and everything else is computed using `Nonlin`$^+$. Note that, in both examples, the nonlinearities involved satisfy the nonlinearity conditions of Theorem A.7. One can easily see this is true in general for the entire subprogram $\pi'$. By induction on $t$, we prove the following claim, which would imply Eq. (36).

**Claim J.1.** For each $v^i, i \geq 1$, the associated random variable $Z^{v^i}$ is zero-mean and 1) independent from $\mathcal{X}^{i-1} \stackrel{\text{def}}{=} \{Z^x : x \text{ introduced before or is } v^{i-1}\}$ if $A^i$ is not diagonal or 2) uncorrelated from $\mathcal{X}^{i-1}$ and $\hat{\mathcal{X}}^{i-1} \stackrel{\text{def}}{=} \{\hat{Z}^x : \text{G-var } x \text{ introduced before or is } v^{i-1}\}$ if $A^i$ is diagonal.

Here, "uncorrelated" means $\mathbb{E}\, Z^{v^i} R = 0$ for all $R \in \mathcal{X}^{i-1} \cup \hat{\mathcal{X}}^{i-1}$. Applying this claim to $v^t$, we get

$$\mathop{\mathbb{E}}_{v,u^1,\ldots,u^t} v^\top (A^t - u^{t\top} A^t u^t) \cdots (A^1 - u^{1\top} A^1 u^1) v = v^\top v^t \xrightarrow{\text{a.s.}} \mathbb{E}\, Z^v Z^{v^t} = 0$$

by Theorem A.7[44] , as desired, assuming rank stability (Assumption E.7). We shall check rank stability after proving this claim.

*Proof.* We proceed by induction on $t$, starting with $t = 1$. Our key tool is Lemma I.6.

**BaseCase:** By the cyclic property of trace, we can WLOG suppose $A^1$ is a polynomial in $\{W, W^\top\}$ for some $W \in \mathcal{W}$. Furthermore, we can assume WLOG that $A^1$ is a monomial in $W$ and $W^\top$. Then $\frac{1}{n} u^{1\top} A^1 u^1 \xrightarrow{\text{a.s.}} \tau^1 \stackrel{\text{def}}{=} \lim_{n\to\infty} \frac{1}{n} \operatorname{tr} A^1$. by something? By Lemma I.6, $Z^{v^1} = Z^{A^1 v} = \tau^1 Z^v + S$ where $S$ is zero-mean and independent from $\mathcal{X}^0$. Then $Z^{v^1} = Z^{A^1 v - (\frac{1}{n} u^{1\top} A^1 u^1) v} = S$ satisfies the required zero-mean and independence property.

**Induction:**

**NondiagonalCase** By IH, $Z^{v^{i-1}}$ is uncorrelated from $\mathcal{X}^{i-2} \cup \hat{\mathcal{X}}^{i-2}$. So we can apply Lemma I.6 in the same fashion as in the base case to obtain the desired result.

**DiagonalCase:** Suppose $A^i = \mathrm{Diag}(x)$ for some $x \in \pi$. By the non-consecutive assumption, $A^{i-1}$ is not diagonal. So IH tells us $Z^{v^{i-1}}$ is independent from $\mathcal{X}^{i-2}$. One can see easily that $Z^{v^i} = Z^{v^{i-1}}(Z^x - \mathbb{E}\, Z^x)$. We need to prove $\mathbb{E}\, Z^{v^i} R = 0$ for all $R \in \mathcal{X}^{i-1} \cup \hat{\mathcal{X}}^{i-1}$. We divide into cases:

1. Suppose $R \in \mathcal{X}^{i-2} \cup \hat{\mathcal{X}}^{i-2}$. Then $\mathbb{E}\, Z^{v^i} R = \mathbb{E}\, Z^{v^{i-1}}(Z^x - \mathbb{E}\, Z^x) R = \mathbb{E}\, Z^{v^{i-1}}\, \mathbb{E}(Z^x - \mathbb{E}\, Z^x) R = 0$ because $Z^{v^{i-1}}$ is independent from $Z^x$ and $R$ and is zero-mean.

[44] While Theorem A.7 is stated for a program (and not a subprogram), its proof can be readily adapted to the subprogram case.

2. Suppose $R \in \hat{\mathcal{X}}^{i-1} \setminus \hat{\mathcal{X}}^{i-2}$. Then $R$ is possibly correlated with $Z^{v^{i-1}}$ but $\{R, Z^{v^{i-1}}\}$ is independent from $\mathcal{X}^{i-2} \ni x$ by Lemma I.5. Therefore, $\mathbb{E}\, Z^{v^i} R = \mathbb{E}\, Z^{v^{i-1}}(Z^x - \mathbb{E}\, Z^x)R = \mathbb{E}\, Z^{v^{i-1}} R\, \mathbb{E}(Z^x - \mathbb{E}\, Z^x) = 0$.
3. Suppose $R \in \mathcal{X}^{i-1} \setminus \mathcal{X}^{i-2}$. Then $R$ decomposes into $R = \tau Z^{v^{i-2}} + S$ by Lemma I.6 for some $\tau \in \mathbb{R}$ and $S$ zero-mean and independent from $\mathcal{X}^{i-2}$. Then

$$\begin{aligned}\mathbb{E}\, Z^{v^i} R = \mathbb{E}\, Z^{v^{i-1}} &= \mathbb{E}\, Z^{v^{i-1}}(Z^x - \mathbb{E}\, Z^x)R \\ &= \mathbb{E}\, Z^{v^{i-1}}(Z^x - \mathbb{E}\, Z^x)(\tau Z^{v^{i-2}} + S) \\ &= \tau\, \mathbb{E}\, Z^{v^{i-1}}(Z^x - \mathbb{E}\, Z^x) Z^{v^{i-2}} + \mathbb{E}\, Z^{v^{i-1}}(Z^x - \mathbb{E}\, Z^x) S \\ &= 0.\end{aligned}$$

   Here $\mathbb{E}\, Z^{v^{i-1}}(Z^x - \mathbb{E}\, Z^x)S = \mathbb{E}\, Z^{v^{i-1}} S\, \mathbb{E}(Z^x - \mathbb{E}\, Z^x) = 0$ because $\{Z^{v^{i-1}}, S\}$ is independent from $Z^x$. Similarly, $\mathbb{E}\, Z^{v^{i-1}}(Z^x - \mathbb{E}\, Z^x)Z^{v^{i-2}} = \mathbb{E}\, Z^{v^{i-1}}\, \mathbb{E}(Z^x - \mathbb{E}\, Z^x)Z^{v^{i-2}} = 0$ because by IH $Z^{v^{i-1}}$ is zero-mean and independent from $\mathcal{X}^{i-2} \ni Z^x, Z^{v^{i-2}}$.

□

Finally, to see rank stability (Assumption E.7), we simply note two points: 1) The subprogram $\pi$ is a NETSOR⊤ program and thus it has rank stability automatically (Remark E.14). 2) $\hat{Z}^x$ has nonzero variance for every G-var $x$ computed in $\pi'$. Since each vector used in `MatMul` in $\pi'$ depends linearly on some unique $\hat{Z}^x$, this shows that the rank in Assumption E.7 is always full, and thus rank stability holds.

□

# K Mathematical Tools

We will use the following trivial but useful fact repeatedly.

**Lemma K.1.** *For an integer $m$, and complex numbers $a_i \in \mathbb{C}$, $i \in [k]$,*

$$\left|\sum_{i=1}^k a_i\right|^m \le k^{m-1} \sum_{i=1}^k |a_i|^m .$$

*Proof.* Expand the power in the LHS using the multinomial theorem, apply AM-GM to each summand, and finally aggregate using triangle inequality. □

## K.1 Probability Facts

**Notations** Given two random variables $X, Z$, and a $\sigma$-algebra $\mathcal{A}$, the notation $X \overset{\mathrm{d}}{=}_{\mathcal{A}} Z$ means that for any integrable function $\phi$ and for any random varible $Z$ measurable on $\mathcal{A}$, $\mathbb{E}\,\phi(X)Z = \mathbb{E}\,\phi(Z)Z$. We say that $X$ is distributed as (or is equal in distribution to) $Z$ conditional on $\mathcal{A}$. In case $\mathcal{A}$ is the trivial $\sigma$-algebra, we just write $X \overset{\mathrm{d}}{=} Z$. If $X$ and $Z$ agrees almost surely, then we write $X \overset{\mathrm{a.s.}}{=} Z$. The expression $X \xrightarrow{\mathrm{d}} Z$ (resp. $X \xrightarrow{\mathrm{a.s.}} Z$) means $X$ converges to $Z$ in distribution (resp. almost surely).

**Lemma K.2.** *Let $\{X_n\}_{n\ge1}$ be a sequence of random variables with zero mean. If for some $p \in \mathbb{N}$ and for all $n$, $\mathbb{E}\, X_n^{2p} \le cn^{-1-\lambda}$, for some $\lambda > 0$, then $X_n \to 0$ almost surely.*

*Proof.* By Markov's inequality, for any $\epsilon > 0$,

$$\begin{aligned}\Pr(|X_n| > \epsilon) = \Pr(X_n^{2p} > \epsilon^{2p}) \le \mathbb{E}\, X_n^{2p}/\epsilon^{2p} \le cn^{-1-\lambda}/\epsilon^{2p} \\ \sum_n \Pr(|X_n| > \epsilon) \le \sum_n cn^{-1-\lambda}/\epsilon^{2p} < \infty.\end{aligned}$$

By Borel-Cantelli Lemma, almost surely, $|X_n| \leq \epsilon$ for all large $n$. Then, if we pick a sequence $\{\epsilon_k > 0\}_k$ converging to 0, we have that, almost surely, for each $k$, $|X_n| \leq \epsilon_k$ for large enough $n$ — i.e. almost surely, $X_n \to 0$. $\square$

The following is a standard fact about multivariate Gaussian conditioning

**Proposition K.3.** *Suppose $\mathbb{R}^{n_1+n_2} \ni x \sim \mathcal{N}(\mu, K)$, where we partition $x = (x_1, x_2) \in \mathbb{R}^{n_1} \times \mathbb{R}^{n_2}$, $\mu = (\mu_1, \mu_2) \in \mathbb{R}^{n_1} \times \mathbb{R}^{n_2}$, and $K = \begin{pmatrix} K_{11} & K_{12} \\ K_{21} & K_{22} \end{pmatrix}$. Then $x_1 \stackrel{\mathrm{d}}{=}_{x_2} \mathcal{N}(\mu|_{x_2}, K|_{x_2})$ where*

$$\begin{aligned} \mu|_{x_2} &= \mu_1 + K_{12} K_{22}^{+} (x_2 - \mu_2) \\ K|_{x_2} &= K_{11} - K_{12} K_{22}^{+} K_{21}. \end{aligned}$$

**Lemma K.4** (Stein's Lemma)**.** *For jointly Gaussian random variables $Z_1, Z_2$ with means $\overline{Z}_1, \overline{Z}_2$, and any differentiable function $\phi : \mathbb{R} \to \mathbb{R}$ where both $\mathbb{E}\, \phi'(Z_1)$ and $\mathbb{E}\, Z_1 \phi(Z_2)$ exist, we have*

$$\mathbb{E}(Z_1 - \overline{Z}_1)\phi(Z_2) = \mathrm{Cov}(Z_1, Z_2)\, \mathbb{E}\, \phi'(Z_2).$$

*More generally, for jointly Gaussian random variables $Z_1, \ldots, Z_k$ with means $\overline{Z}_1, \ldots, \overline{Z}_k$, and any differentiable function $\phi : \mathbb{R}^k \to \mathbb{R}$, we have*

$$\mathbb{E}(Z_1 - \overline{Z}_1)\phi(Z_1, \ldots, Z_k) = \sum_{j=1}^{k} \mathrm{Cov}(Z_1, Z_j)\, \mathbb{E}\, \partial_j \phi(Z_1, \ldots, Z_k)$$

*whenever both sides are finite.*

**Lemma K.5.** *Let $X = (X_1, \ldots, X_k) \in \mathbb{R}^k$ be a multivariate Gaussian with 0 mean and nondegenerate covariance $\Omega$. Let $X_{\bar{i}}$ denote the vector $(X_1, \ldots, X_{i-1}, X_{i+1}, \ldots, X_k) \in \mathbb{R}^{k-1}$. Likewise, let $\Omega_{i\bar{i}} \in \mathbb{R}^{k-1}$ (resp. $\Omega_{\bar{i}i} \in \mathbb{R}^{k-1}$) be the $i$th row (resp. column) with entry $i$ removed, and $\Omega_{\bar{i}\bar{i}} \in \mathbb{R}^{(k-1)\times(k-1)}$ be the submatrix of $\Omega$ obtained by removing the $i$th row and $i$th column. Then for any $i \in [k]$ and any $f : \mathbb{R}^k \to \mathbb{R}$,*

$$\mathbb{E}\, X_i f(X) = \sum_{j=1}^{k} a_j \Omega_{ji}$$

*whenever both sides are defined, where*

$$a_j \stackrel{\text{def}}{=} \frac{\mathbb{E}(X_j - \mathbb{E}[X_j \mid X_{\bar{j}}]) f(X)}{\mathrm{Var}(X_j \mid X_{\bar{j}})} = \frac{\mathbb{E}(X_j - \Omega_{j\bar{j}} \Omega_{\bar{j}\bar{j}}^{+} X_{\bar{j}}) f(X)}{\Omega_{jj} - \Omega_{j\bar{j}} \Omega_{\bar{j}\bar{j}}^{+} \Omega_{\bar{j}j}}.$$

*Note that, for any $j$, $a_j$ does not depend on $i$.*

*Proof.* By a standard density argument, we may assume $f$ is differentiable. Then by multivariate Stein's Lemma (Lemma K.4),

$$\mathbb{E}\, X_i f(X) = \sum_{j=1}^{k} \Omega_{ij}\, \mathbb{E}\, \partial_j f(X).$$

By applying bivariate Stein's Lemma (Lemma K.4) to conditional distributions, we get

$$\mathbb{E}\, \partial_j f(X) = \mathop{\mathbb{E}}_{X_{\bar{j}}} \mathop{\mathbb{E}}_{X_j | X_{\bar{j}}} \partial_j f(X) = \mathop{\mathbb{E}}_{X_{\bar{j}}} \mathop{\mathbb{E}}_{X_j | X_{\bar{j}}} \frac{(X_j - \mathbb{E}[X_j \mid X_{\bar{j}}]) f(X)}{\mathrm{Var}(X_j \mid X_{\bar{j}})} = \mathop{\mathbb{E}}_{X} \frac{(X_j - \mathbb{E}[X_j \mid X_{\bar{j}}]) f(X)}{\mathrm{Var}(X_j \mid X_{\bar{j}})}.$$

The other equality follows from straightforward calculations. $\square$

**Lemma K.6.** *Let $\Phi : \mathbb{R}^n \to \mathbb{R}$ be measurable. Then for $z \sim \mathcal{N}(\zeta, \Sigma)$, the following Hessian and gradient matrices are equal:*

$$\frac{\mathrm{d}^2}{\mathrm{d}\zeta^2}\, \mathbb{E}\, \Phi(z) = 2 \frac{\mathrm{d}}{\mathrm{d}\Sigma}\, \mathbb{E}\, \Phi(z)$$

*whenever both sides exist.*

*Proof.* First assume $\Sigma$ is invertible. We check

$$\begin{aligned}
\frac{\mathrm{d}}{\mathrm{d}\zeta} e^{-\frac{1}{2}(\zeta-z)\Sigma^{-1}(\zeta-z)} &= -\Sigma^{-1}(\zeta - z) e^{-\frac{1}{2}(\zeta-z)\Sigma^{-1}(\zeta-z)} \\
\frac{\mathrm{d}^2}{\mathrm{d}\zeta^2} e^{-\frac{1}{2}(\zeta-z)\Sigma^{-1}(\zeta-z)} &= \left[-\Sigma^{-1} + \Sigma^{-1}(\zeta - z)(\zeta - z)^\top \Sigma^{-1}\right] e^{-\frac{1}{2}(\zeta-z)\Sigma^{-1}(\zeta-z)} \\
\frac{\mathrm{d}}{\mathrm{d}\Sigma} \frac{e^{-\frac{1}{2}(\zeta-z)\Sigma^{-1}(\zeta-z)}}{\det(2\pi\Sigma)^{1/2}} &= \frac{1}{2}\left[-\Sigma^{-1} + \Sigma^{-1}(\zeta - z)(\zeta - z)^\top \Sigma^{-1}\right] \frac{e^{-\frac{1}{2}(\zeta-z)\Sigma^{-1}(\zeta-z)}}{\det(2\pi\Sigma)^{1/2}} \\
&= \frac{1}{2}\frac{\mathrm{d}^2}{\mathrm{d}\zeta^2} e^{-\frac{1}{2}(\zeta-z)\Sigma^{-1}(\zeta-z)}.
\end{aligned}$$

Integrating against $\Phi$ gives the result. For general $\Sigma$, apply a continuity argument, since the set of invertible $\Sigma$s is dense inside the set of all PSD $\Sigma$. ☐

### K.2 Gaussian Conditioning Trick

**Review of Moore-Penrose Pseudoinverse** We first recall Moore-Penrose pseudoinverse and some properties of it.

**Definition K.7.** For $A \in \mathbb{R}^{n\times m}$, a pseudoinverse of $A$ is defined as a matrix $A^+ \in \mathbb{R}^{m\times n}$ that satisfies all of the following criteria

$$AA^+A = A, \qquad A^+AA^+ = A^+, \qquad (AA^+)^\top = AA^+, \qquad (A^+A)^\top = A^+A.$$

The following facts are standard.

- If $A$ has real entries, then so does $A^+$.
- The pseudoinverse always exists and is unique.
- When $A$ is invertible, $A^+ = A^{-1}$.
- $(A^\top)^+ = (A^+)^\top$, which we denote as $A^{+\top}$.
- $A^+ = (A^\top A)^+ A^\top = A^\top(AA^\top)^+$.
- $AA^+$ is the orthogonal projector to the column space of $A$; $I - A^+A$ is the orthogonal project to the null space of $A$.
- If $A$ has singular value decomposition $A = U\Lambda V$ where $U$ and $V$ are orthogonal and $\Lambda$ has the singular values on its diagonal, then $A^+ = V^\top\Lambda^+U^\top$ where $\Lambda^+$ inverts all nonzero entries of $\Lambda$.
- For any collection of vectors $\{v_i\}_{i=1}^n$ in a Hilbert space, $w \mapsto \sum_{i,j=1}^n v_i(\Sigma^+)_{ij}\langle v_j, w\rangle$, where $\Sigma_{ij} = \langle v_i, v_j\rangle$, is the projection operator to the linear span of $\{v_i\}_{i=1}^n$.

We present a slightly more general versions of lemmas from Bayati and Montanari [6] that deal with singular matrices.

**Lemma K.8.** *Let $z \in \mathbb{R}^n$ be a random vector with i.i.d. $\mathcal{N}(0, v^2)$ entries and let $D \in \mathbb{R}^{m\times n}$ be a linear operator. Then for any constant vector $b \in \mathbb{R}^n$ the distribution of $z$ conditioned on $Dz = b$ satisfies:*

$$z \overset{\mathrm{d}}{=}_{Dz=b} D^+b + \Pi\tilde{z}$$

*where $D^+$ is the (Moore-Penrose) pseudoinverse, $\Pi$ is the orthogonal projection onto subspace $\{z : Dz = 0\}$, and $\tilde{z}$ is a random vector of i.i.d. $\mathcal{N}(0, v^2)$.*

*Proof.* When $D = [I_{m\times m}|0_{m\times n-m}]$, this claim is immediate. By rotational symmetry, this shows that, for any vector space $\mathcal{V}$ and $v$ orthogonal to it, conditioning $z$ on $\mathcal{V} + v$ yields a Gaussian centered on $v$ with covariance determined by $\Pi_{\mathcal{V}} z$. Then the lemma in the general case is implied by noting that $\{z : Dz = b\}$ can be decomposed as $\{z : Dz = 0\} + D^+b$. ☐

**Lemma K.9.** *Let $A \in \mathbb{R}^{n\times m}$ be a matrix with random Gaussian entries, $A_{ij} \sim \mathcal{N}(0, \sigma^2)$. Consider fixed matrices $Q \in \mathbb{R}^{m\times q}, Z \in \mathbb{R}^{n\times q}, P \in \mathbb{R}^{n\times p}, X \in \mathbb{R}^{m\times p}$. Suppose there exists a solution in $A$ to the equations $Z = AQ$ and $X = A^\top P$. Then the distribution of $A$ conditioned on $Z = AQ$ and $X = A^\top P$ is*

$$A \overset{\mathrm{d}}{=}_{Z=AQ,X=A^\top P} E + \Pi_P^\perp \tilde{A} \Pi_Q^\perp$$

*where*

$$E = ZQ^+ + P^{+\top} X^\top - P^{+\top} P^\top Y Q^+,$$

*$\tilde{A}$ is an iid copy of $A$, and $\Pi_P^\perp = I - \Pi_P$ and $\Pi_Q^\perp = I - \Pi_Q$ in which $\Pi_P = PP^+$ and $\Pi_Q = QQ^+$ are the orthogonal projection to the space spanned by the column spaces of $P$ and $Q$ respectively.*

*Proof.* We apply Lemma K.8 to $D : A \mapsto (AQ, P^\top A)$. The pseudoinverse of $D$ applied to $(Z, X^\top)$ can be formulated as the unique solution of

$$\operatorname*{argmin}_A \left\{ \|A\|_F^2 : AQ = Z, P^\top A = X^\top \right\}$$

where $\| - \|_F$ denotes Frobenius norm. We check that $E$ is a 1) a solution to $AQ = Z, P^\top A = X^\top$ and 2) the minimal norm solution.

We have $EQ = ZQ^+Q + P^{+\top} X^\top Q - P^{+\top} P^\top Y Q^+ Q$. Note that $YQ^+Q = Z$ because $Z = AQ \implies YQ^+Q = AQQ^+Q = AQ = Z$. So $EQ = Z + P^{+T}(X^\top Q - P^\top Z)$. But $X^\top Q = P^\top AQ = P^\top Z$, so $EQ = Z$ as desired. A similar, but easier reasoning, gives $P^\top E = X^\top$. This verifies that $E$ is a solution.

To check that $E$ is minimal norm, we show that it satisfies the stationarity of the Lagrangian

$$L(A, \Theta, \Gamma) = \|A\|_F^2 + \langle \Theta, Z - AQ \rangle + \langle \Gamma, X - A^\top P \rangle.$$

So $\frac{\partial L}{\partial A} = 0 \implies 2A = \Theta Q^\top + P\Gamma^\top$ for some choices of $\Theta \in \mathbb{R}^{n\times q}$ and $\Gamma \in \mathbb{R}^{m\times p}$. For $\Theta = 2Z(Q^\top Q)^+$ and $\Gamma^\top = 2(P^\top P)^+[X^\top - P^\top Z Q^\top]$, we can check that

$$\begin{aligned}
\Theta Q^\top + P\Gamma^\top &= 2Z(Q^\top Q)^+ Q^\top + 2P(P^\top P)^+ [X^\top - P^\top Z Q^+] \\
&= 2ZQ^+ + 2P^{+\top} X^\top - 2P^{+\top} P^\top Z Q^+ \\
&= 2E
\end{aligned}$$

as desired. ☐

### K.3 Hermite Polynomials

We follow a presentation roughly given by O'Donnell [33].

**Definition K.10.** Let $\mathrm{He}_n(x)$ be the *probabilist's Hermite polynomial*, given by the generating function $e^{xt - \frac{1}{2}t^2} = \sum_{n=0}^\infty \mathrm{He}_n(x) \frac{t^n}{n!}$. Let $L^2(\mathbb{R}; \mathcal{N}(0,1))$ be the space of square-integrable functions against the standard Gaussian measure, equipped with inner product $\langle \phi, \psi \rangle_G = \mathbb{E}_{x\sim\mathcal{N}(0,1)} \phi(x)\psi(x)$ and norm $\|\phi\|_G^2 = \langle \phi, \phi \rangle_G$. Let $H_n(x) = \mathrm{He}_n(x)/\|\mathrm{He}_n\|_G$ be the normalized versions.

**Fact K.11.** $\{\mathrm{He}_n(x)\}_{n\geq 0}$ *form an orthogonal basis for* $L^2(\mathbb{R}; \mathcal{N}(0,1))$ *and* $\{H_n(x)\}_{n\geq 0}$ *form an orthonormal basis for* $L^2(\mathbb{R}; \mathcal{N}(0,1))$.

**Fact K.12.** $\|\mathrm{He}_n\|_G^2 = n!$ *so that* $H_n(x) = \mathrm{He}_n(x)/\sqrt{n!}$.

**Fact K.13.** *Let $\phi, \psi : \mathbb{R} \to \mathbb{R}$ be square integrable against $\mathcal{N}(0,1)$. Suppose we have the expansions in the orthonormal Hermite basis*

$$\phi(x) = a_0 H_0(x) + a_1 H_1(x) + \cdots, \quad \psi(x) = b_0 H_0(x) + b_1 H_1(x) + \cdots.$$

*Let $(z_1, z_2) \sim \mathcal{N}(0, C)$ where $C_{11} = C_{22} = 1$ and $C_{12} = \rho \in [-1, 1]$. Then we have the absolutely convergence series*

$$\mathbb{E}\,\phi(z_1)\psi(z_2) = a_0 b_0 + a_1 b_1 \rho + a_2 b_2 \rho^2 + \cdots.$$

Suppose $u^1, \ldots, u^k$ are unit vectors in $\mathbb{R}^k$, and let $\lambda_{ij} \stackrel{\text{def}}{=} \langle u^i, u^j \rangle$. Construct a zero mean Gaussian vector $z = (z_1, \ldots, z_k)$ such that $\mathbb{E}\, z_i z_j = \lambda_{ij}$. Note that $z \stackrel{\text{d}}{=} Ug$ where $g = (g_1, \ldots, g_k)$ is a standard Gaussian vector and $U = (u^i_j)_{i,j=1}^k$ is the matrix with $u^i$ as rows. Then for any $s = (s_1, \ldots, s_k)$ we can compute

$$\begin{aligned}
\mathbb{E}\exp(\langle s, z\rangle) &= \mathbb{E}\exp(s^\top U g) = \mathbb{E}\prod_i \exp(g_i (U^\top s)_i) \\
&= \prod_i \mathbb{E}\exp(g_i(U^\top s)_i) && \text{by independence of } \{g_i\}_i \\
&= \prod_i \exp\left(\frac{1}{2}(U^\top s)_i^2\right) = \exp\left(\frac{1}{2}\sum_i (U^\top s)_i^2\right) \\
&= \exp\left(\frac{1}{2}\|U^\top s\|^2\right) = \exp\left(\frac{1}{2}\sum_{i,j}\langle u^i, u^j\rangle s_i s_j\right) \\
&= \exp\left(\frac{1}{2}\sum_{i,j}\lambda_{ij} s_i s_j\right).
\end{aligned}$$

Dividing by $\exp\left(\frac{1}{2}\sum_i s_i^2\right)$, we obtain

$$\begin{aligned}
\mathbb{E}\exp\left(\sum_i s_i z_i - s_i^2\right) &= \exp\left(\sum_{i<j}\lambda_{ij} s_i s_j\right) \\
\mathbb{E}\prod_i \sum_m \mathrm{He}_m(z_i)(m!)^{-1} s_i^m &= \prod_{i<j}\sum_n (n!)^{-1}\left(\lambda_{ij} s_i s_j\right)^n \\
\sum_{(m_i)_{i=1}^k} \prod_i \frac{s_i^{m_i}}{m_i!}\,\mathbb{E}\prod_i \mathrm{He}_{m_i}(z_i) &= \sum_{(n_{(ij)})_{i<j}} \prod_i s_i^{\sum_{j\neq i} n_{(ij)}} \prod_{i<j} \frac{\lambda_{ij}^{n_{(ij)}}}{n_{(ij)}!}
\end{aligned}$$

where $m_i \geq 0$ for all $i$, and $n_{(ij)} = n_{(ji)} \geq 0$ are indexed by unordered sets $\{i, j\}$. Matching coefficients of $s$, we get

**Theorem K.14.** *For any sequence $(m_i \geq 0)_{i=1}^k$,*

$$\mathbb{E}\prod_i \mathrm{He}_{m_i}(z_i) = \left(\prod_r m_r!\right)\left(\prod_{i<j}\frac{\lambda_{ij}^{n_{(ij)}}}{n_{(ij)}!}\right)$$

$$\mathbb{E}\prod_i H_{m_i}(z_i) = \left(\prod_r \sqrt{m_r!}\right)\left(\prod_{i<j}\frac{\lambda_{ij}^{n_{(ij)}}}{n_{(ij)}!}\right)$$

*whenever there are $(n_{(ij)} \geq 0)_{i<j}$ such that, for all $i$, $m_i = \sum_{j\neq i} n_{(ij)}$. $\mathbb{E}\prod_i \mathrm{He}_{m_i}(z_i) = 0$ otherwise.*

In particular,

**Theorem K.15.** *If $\phi_i : \mathbb{R} \to \mathbb{R}$ has Hermite expansion $\phi_i(z) = \sum_{u=0}^\infty a_{iu} H_u(z) = \sum_{u=0}^\infty b_{iu}\mathrm{He}_u(z)$ where $b_{iu} = a_{iu}/\sqrt{u!}$, then*

$$\begin{aligned}
\mathbb{E}\prod_i \phi_i(z_i) &= \sum_{(n_{(ij)})_{i<j}} \left(\prod_r b_{r m_r} m_r!\right)\left(\prod_{i<j}\frac{\lambda_{ij}^{n_{(ij)}}}{n_{(ij)}!}\right) \\
&= \sum_{(n_{(ij)})_{i<j}} \left(\prod_r a_{r m_r}\sqrt{m_r!}\right)\left(\prod_{i<j}\frac{\lambda_{ij}^{n_{(ij)}}}{n_{(ij)}!}\right) \\
&= \sum_{(n_{(ij)})_{i<j}} \left(\prod_r a_{r m_r}\sqrt{\binom{m_i}{\{n_{(ij)}\}_{j\neq i}}}\right)\left(\prod_{i<j}\lambda_{ij}^{n_{(ij)}}\right)
\end{aligned}$$

*where $m_i = \sum_{j \neq i} n_{(ij)}$, whenever the RHS is absolutely convergent.*

**Lemma K.16.** *Suppose $\phi_i, i \in [k]$ are as in Theorem K.15, with additionally the constraint that we have an index set $I \subseteq [k]$ such that $b_{i0} = a_{i0} = 0$ (i.e. $\mathbb{E}\, \phi_i(z_i) = 0$) for all $i \in I$. Assume that, for some $\lambda < 1/2$, $|\lambda_{ij}| \leq \lambda/(k-1)$ for all $i \neq j$. Then*

$$\left| \mathbb{E} \prod_{i=1}^{k} \phi_i(z_i) \right| \leq C_{k,|I|} \left( \prod_{r=1}^{k} \|\phi_r\|_G \right) \lambda^{\lceil |I|/2 \rceil}$$

*for some constant $C_{k,|I|}$ depending on $k$ and $|I|$ but independent of $\{\phi_i\}_i$ and $\lambda$.*

*Proof.* In the notation of Theorem K.15, $\binom{m_i}{\{n_{(ij)}\}_{j \neq i}} \leq (k-1)^{m_i}$ by the multinomial theorem. Thus

$$\begin{aligned}
\left| \mathbb{E} \prod_{i=1}^{k} \phi_i(z_i) \right| &\leq \sum_{\substack{(n_{(ij)})_{i<j}: \\ \forall r \in I, m_r \geq 1}} \left| \left( \prod_{r=1}^{k} a_{r m_r} \sqrt{\binom{m_r}{\{n_{(rj)}\}_{j \neq r}}} \right) \left( \prod_{i<j} \lambda_{ij}^{n_{(ij)}} \right) \right| \\
&\leq \sum_{\substack{(n_{(ij)})_{i<j}: \\ \forall r \in I, m_r \geq 1}} \left( \prod_{r=1}^{k} \|\phi_r\|_G \sqrt{(k-1)^{m_r}} \right) \left( \prod_{i<j} \left( \frac{\lambda}{k-1} \right)^{n_{(ij)}} \right) \\
&\leq \sum_{\substack{(n_{(ij)})_{i<j}: \\ \forall r \in I, m_r \geq 1}} \left( \prod_{r=1}^{k} \|\phi_r\|_G \right) \lambda^{\sum_{i<j} n_{(ij)}} \\
&= \left( \prod_{r=1}^{k} \|\phi_r\|_G \right) \left( B_{|I|} \lambda^{\lceil |I|/2 \rceil} (1 + o(1)) \right).
\end{aligned}$$

where $B_V$ is the number of ways to cover $V$ vertices with $\lceil V/2 \rceil$ edges, and $o(1)$ is a term that goes to 0 as $\lambda \to 0$ and is bounded above by a function of $k$ whenever $\lambda < 1/2$. Then an appropriate $C_{k,|I|}$ can be chosen to obtain the desired result. $\square$

**Lemma K.17.** *Suppose $\phi_i, i \in [k]$ are as in Theorem K.15, with additionally the constraint that, we have some index set $I \subseteq [3, k]$ such that for all $i \in I$, $b_{i0} = a_{i0} = 0$ (i.e. $\mathbb{E}\, \phi_i(z_i) = 0$). Assume that $|\lambda_{12}| \leq 1/2$, for some $\lambda < 1/\sqrt{8}$, $|\lambda_{ij}| \leq \lambda/(k-1)$ for all $i \neq j$ and $\{i, j\} \neq \{1, 2\}$. Then*

$$\left| \mathbb{E} \prod_{i=1}^{k} \phi_i(z_i) \right| \leq C'_{k,|I|} \left( \prod_{r=1}^{k} \|\phi_r\|_G \right) \lambda^{\lceil |I|/2 \rceil}$$

*for some constant $C'_k$ depending on $k$ and $I$ but independent of $\{\phi_i\}_i$ and $\lambda$.*

*Proof.* Define $\mathcal{P} = \{(i,j) : 1 \neq i < j \neq 2\}$ and $\mathcal{Q} = \{(i,j) : i < j \text{ and } (i = 1 \text{ XOR } j = 2)\}$. Also write $R = \prod_{r=1}^k \|\phi_r\|_G$. As in the above proof,

$$
\begin{aligned}
&|\mathbb{E}\prod_{i=1}^k \phi_i(z_i)| \\
&\leq \sum_{\substack{(n_{(ij)})_{i<j}:\\ \forall r \in I, m_r \geq 1}} \left| \left(\prod_{r=1}^k a_{rm_r}\sqrt{\binom{m_r}{\{n_{(rj)}\}_{j\neq r}}}\right)\left(\prod_{(i,j)\in\mathcal{P}} \lambda_{ij}^{n_{(ij)}}\right) 2^{-n_{(12)}} \right| \\
&\leq \sum_{\substack{(n_{(ij)})_{i<j}:\\ \forall r \in I, m_r \geq 1}} R\sqrt{\prod_{r=1}^2 \binom{m_r}{n_{(12)}}\binom{m_r - n_{(12)}}{\{n_{(rj)}\}_{j\notin\{1,2\}}}}\prod_{r=3}^k\sqrt{(k-1)^{m_r}}\left(\prod_{(i,j)\in\mathcal{P}}\left(\frac{\lambda}{k-1}\right)^{n_{(ij)}}\right)2^{-n_{(12)}} \\
&\leq R\sum_{\substack{(n_{(ij)})_{i<j}:\\ \forall r \in I, m_r \geq 1}} \sqrt{\prod_{r=1}^2\binom{m_r}{n_{(12)}}(k-1)^{m_r - n_{(12)}}}(k-1)^{\frac{1}{2}\sum_{r=3}^k m_r}\left(\frac{\lambda}{k-1}\right)^{\sum_{i<j} n_{(ij)} - n_{(12)}} 2^{-n_{(12)}} \\
&= R\sum_{\substack{(n_{(ij)})_{i<j}:\\ \forall r \in I, m_r \geq 1}} \sqrt{\prod_{r=1}^2\binom{m_r}{n_{(12)}}}\lambda^{\sum_{i<j} n_{(ij)} - n_{(12)}}2^{-n_{(12)}} \\
&\leq R\sum_{\substack{(n_{(ij)})_{i<j}:\\ \forall r \in I, m_r \geq 1}} \binom{\frac{m_1+m_2}{2}}{n_{(12)}}2^{-n_{(12)}}\lambda^{\sum_{i<j} n_{(ij)} - n_{(12)}} \\
&\leq R\sum_{\substack{(n_{(ij)})_{i<j}:\\ \forall r \in I, m_r \geq 1}} \binom{n_{(12)} + \frac{1}{2}m_{(12)}}{n_{(12)}}2^{-n_{(12)}}\lambda^{\sum_{i<j} n_{(ij)} - n_{(12)}} \\
&\qquad \text{where } m_{(12)} = \sum_{(i,j)\in\mathcal{Q}} n_{(ij)} \\
&\leq 2R\sum_{\substack{(n_{(ij)})_{(i,j)\in\mathcal{P}}:\\ \forall r \in I, m_r \geq 1}} \left(\frac{1}{1-1/2}\right)^{1+\frac{1}{2}m_{(12)}}\lambda^{\sum_{(i,j)\in\mathcal{P}} n_{(ij)}} \\
&\leq 2R\sum_{\substack{(n_{(ij)})_{(i,j)\in\mathcal{P}}:\\ \forall r \in I, m_r \geq 1}} (\sqrt{2}\lambda)^{\sum_{(i,j)\in\mathcal{Q}} n_{(ij)}}\lambda^{\sum_{(i,j)\in\mathcal{P}\setminus\mathcal{Q}} n_{(ij)}} \\
&\leq 2R\left(2^{|I|/4}B_{|I|}\lambda^{\lceil |I|/2\rceil}(1+o(1))\right)
\end{aligned}
$$

where $B_{|I|}$ is the number of ways of covering $|I|$ vertices with $\lceil\frac{|I|}{2}\rceil$ edges, and $o(1)$ is a term that goes to 0 as $\lambda \to 0$ and is upper bounded by a function of $k$ for all $\lambda < 1/\sqrt{8}$. Choosing the appropriate constant $C'_{k,|I|}$ then gives the result. □

### K.4 Bounding the Off-Diagonal Correlations of a Projection Matrix

**Lemma K.18.** *Let $\Pi \in \mathbb{R}^{n\times n}$ be an orthogonal projection matrix. Then each diagonal entry $\Pi_{ii} \in [0,1]$.*

*Proof.* Because $\Pi = \Pi^2$, we have for each $i$, $\Pi_{ii} = \sum_j \Pi_{ij}^2 \implies \Pi_{ii}(1-\Pi_{ii}) = \sum_{j\neq i}\Pi_{ij}^2 \geq 0 \implies \Pi_{ii} \in [0,1]$. □

**Lemma K.19.** *Let $\Pi \in \mathbb{R}^{n\times n}$ be an orthogonal projection matrix of rank $k$. Suppose its diagonal entries are strictly positive. Consider the correlation matrix $C \stackrel{\text{def}}{=} D^{-1/2}\Pi D^{-1/2}$ where $D = \mathrm{Diag}(\Pi)$. Then the off-diagonal entries of $C$ satisfy $\sum_{i<j} C_{ij}^2 \leq 0.5(r^2+r)$, where $r = n-k$.*

*Proof.* Because $\Pi = \Pi^2$, we have for each $i$, $\Pi_{ii} = \sum_j \Pi_{ij}^2 = \sum_j \Pi_{ii}\Pi_{jj}C_{ij}^2 \implies 1 - \Pi_{ii} = \sum_{j\neq i} \Pi_{jj}C_{ij}^2$. At the same time, each $C_{ij}^2 \in [0,1]$. Thus we seek an upper bound on the following linear program in the $n(n-1)/2$ variables $C_{(ij)}^2$ (which identiy $C_{ij} = C_{ji} = C_{(ij)}$).

$$\begin{aligned} \text{Maximize } & \sum_{i\neq j} C_{(ij)}^2 \\ \text{s.t. } & \forall i, 1 - \Pi_{ii} = \sum_{j\neq i} \Pi_{jj} C_{(ij)}^2 \\ & \forall i < j, C_{(ij)}^2 \in [0,1]. \end{aligned}$$

This LP has the dual

$$\begin{aligned} \text{Minimize } & \sum_{i<j} \tau_{ij} + \sum_{i=1}^n (1-\Pi_{ii})\zeta_i \\ \text{s.t. } & \forall i < j, \tau_{ij} + \zeta_i \Pi_{jj} + \zeta_j \Pi_{ii} \geq 1 \\ & \forall i < j, \tau_{ij} \geq 0 \\ & \forall i, \zeta_i \in \mathbb{R}. \end{aligned}$$

Any feasible value of the dual LP is an upper bound on the original LP. We now set the dual variables. We will set $\tau_{ij} = 1 - \zeta_i \Pi_{jj} - \zeta_j \Pi_{ii}$ for all $i < j$, so that the objective is now

$$\binom{n}{2} - (k-1)\sum_{i=1}^n \zeta_i.$$

It remains to 1) set the variables $\zeta_i$ and 2) verify that $1 - \zeta_i \Pi_{jj} + \zeta_j \Pi_{ii} \geq 0$ for all $i < j$.

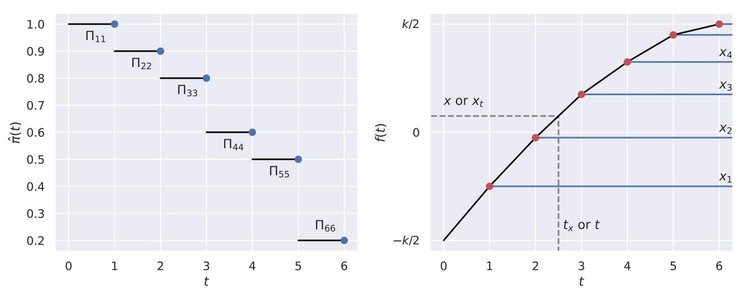


Figure 2: Illustration of $\hat{\pi}$, $x_t$, and $t_x$.

WLOG, assume $\Pi_{11} \geq \cdots \geq \Pi_{nn} > 0$. Define the function $\pi : (0, n] \to \mathbb{R}$ as the piecewise constant extension of the diagonal of $\Pi$ to a real function on $(0, n]$:

$$\pi(t) \stackrel{\text{def}}{=} \Pi_{ii}, \quad \text{where} \quad i = \lceil t \rceil.$$

Note $\pi$ is nonincreasing and takes values in $(0, 1]$. Define $f$ as the integral of $\pi$:

$$f(s) \stackrel{\text{def}}{=} \int_0^s \pi(t)\, \mathrm{d}t.$$

Note $f$ is increasing and concave because $\pi$ is positive and nonincreasing, so $f$ has an inverse $f^{-1}$. Additionally, $f(0) = 0, f(n) = k$.

Let

$$t_x \stackrel{\text{def}}{=} f^{-1}(k/2 + x), \quad x_t \stackrel{\text{def}}{=} f(t) - k/2.$$

Now define $\hat{\pi} : [-k/2, k/2] \to (0, 1]$ by

$$\hat{\pi}(x) = \pi(t_x) = \pi(\lceil t_x \rceil), \quad \text{so that for all } t \in (0, n], \quad \hat{\pi}(x_t) = \pi(t).$$

Fig. 2 illustrates the above definitions. Suppose

$$\{x_1, \ldots, x_n\} \text{ and } \{-x_1, \ldots, -x_n\} \text{ are disjoint and do not contain } 0. \tag{37}$$

We can assume this WLOG as the set of $\Pi$ satisfying this property is dense and the function $\Pi \mapsto \sum C_{ij}^2$ is continuous so our bound applies to all $\Pi$ be continuity. This assumption implies that any line segment $[-\delta, +\delta]$ has at most one endpoint in $\{x_1, \ldots, x_n\}$.

Now we define $\hat{\zeta} : [-k/2, k/2] \to \mathbb{R}$ as the unique function satisfying

$$\hat{\zeta}(x)\hat{\pi}(-x) + \hat{\zeta}(-x)\hat{\pi}(x) = 1. \tag{38}$$

and

$$\hat{\zeta}(x) = \hat{\zeta}(y) \quad \text{if } \lceil t_x \rceil = \lceil t_y \rceil \tag{39}$$

Note Eq. (38) implies $\hat{\zeta}(0) = \frac{1}{2\hat{\pi}(0)}$ and

$$\frac{\hat{\zeta}(x)}{\hat{\pi}(x)} + \frac{\hat{\zeta}(-x)}{\hat{\pi}(-x)} = \frac{1}{\hat{\pi}(x)\hat{\pi}(-x)}. \tag{40}$$

By Eq. (39), $\hat{\zeta}$ is constant on each subinterval of $[-k/2, k/2]$ in the partition induced by $\{x_1, \ldots, x_n\}$. A second of thought shows there is a unique $\hat{\zeta}$ satisfying Eq. (38) and Eq. (39); this is illustrated by by Fig. 3.

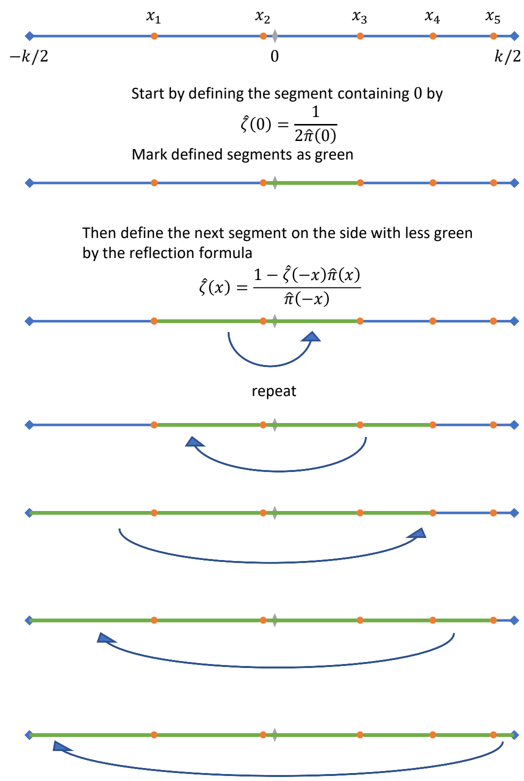


Figure 3: Illustration of how to define $\hat{\zeta}$.

Thus we may set the dual variables

$$\zeta_i \stackrel{\text{def}}{=} \hat{\zeta}(x), \quad \text{where} \quad \lceil t_x \rceil = i.$$

We check this is a valid assignment of dual variables (in that $\tau_{ij} \geq 0$ are satisfied) in Claim K.3.

Let $r = n - k$. It turns out, given Claim K.1 and Claim K.2, we have

$$\sum_{i=1}^{n} \zeta_i \geq \frac{n+r}{2}$$

so that the objective value given the above dual variables is

$$\binom{n}{2} - \frac{n+r}{2}(n-r-1) = \frac{1}{2}(n^2 - n) - \frac{1}{2}(n^2 - r^2 - n - r) = \frac{1}{2}(r^2 + r)$$

as desired.

**Claim K.1.**

$$\sum_{i=1}^{n} \zeta_i = \int_{-k/2}^{k/2} \frac{\hat{\zeta}(x)}{\hat{\pi}(x)}\, \mathrm{d}x = \int_0^{k/2} \frac{\mathrm{d}x}{\hat{\pi}(x)\hat{\pi}(-x)}.$$

*Proof.* The first equality follows from

$$\int_{x : \lceil t_x \rceil = i} \frac{\hat{\zeta}(x)}{\hat{\pi}(x)}\, \mathrm{d}x = \int_{f(i-1)-k/2}^{f(i)-k/2} \frac{\hat{\zeta}(x)}{\hat{\pi}(x)}\, \mathrm{d}x = \int_{f(i-1)}^{f(i)} \frac{\zeta_i}{\pi(i)}\, \mathrm{d}x = (f(i) - f(i-1))\frac{\zeta_i}{\pi(i)} = \zeta_i.$$

The second equality follows from folding the integral around $x = 0$ and applying Eq. (38). $\square$

**Claim K.2.** *With $r = n - k$,*

$$\int_0^{k/2} \frac{\mathrm{d}x}{\hat{\pi}(x)\hat{\pi}(-x)} \geq \frac{n+r}{2}$$

*Proof.* Note $\int_{-k/2}^{k/2} \frac{\mathrm{d}x}{\hat{\pi}(x)} = \sum_{i=1}^n 1 = n$ by same reasoning as in Claim K.1. We also have $\hat{\pi}(x) \leq 1 \implies \frac{1}{\hat{\pi}(x)} \geq 1$ for all $x$. Thus $\int_0^{k/2} \frac{\mathrm{d}x}{\hat{\pi}(x)\hat{\pi}(-x)}$ is at least

$$\frac{1}{2}\inf\left\{\int_{-k/2}^{k/2} g(x)g(-x)\,\mathrm{d}x \mid g : [-k/2, k/2] \to [1,\infty), \int_{-k/2}^{k/2} g(x)\,\mathrm{d}x = n, g \text{ nondecreasing}\right\}.$$

It's not hard to see this infimum is achieved by $g^*(x) = 1$ for all $x \in [-k/2, k/2)$ with a delta jump of mass $n - k = r$ at the point $k/2$. The value for this $g^*(x)$ is $\frac{1}{2}(k + 2r) = \frac{n+r}{2}$. ☐

**Claim K.3.** *For any $x, y \in [-k/2, k/2]$,*

$$\hat{\zeta}(x)\hat{\pi}(y) + \hat{\zeta}(y)\hat{\pi}(x) \leq 1.$$

*Proof.* We show the equivalent claim that

$$\frac{\hat{\zeta}(x)}{\hat{\pi}(x)} + \frac{\hat{\zeta}(y)}{\hat{\pi}(y)} \leq \frac{1}{\hat{\pi}(x)\hat{\pi}(y)}.$$

Suppose $-x < y$, then by telescoping and Claim K.4,

$$\begin{aligned}
\frac{\hat{\zeta}(x)}{\hat{\pi}(x)} + \frac{\hat{\zeta}(y)}{\hat{\pi}(y)} &= \frac{\hat{\zeta}(x)}{\hat{\pi}(x)} + \frac{\hat{\zeta}(-x)}{\hat{\pi}(-x)} + \sum_{i=\lceil t_{-x}\rceil}^{\lceil t_y\rceil}\left(\frac{\zeta_{i+1}}{\pi(i+1)} - \frac{\zeta_i}{\pi(i)}\right) \\
&= \frac{1}{\hat{\pi}(x)\hat{\pi}(-x)} + \sum_{i=\lceil t_{-x}\rceil}^{\lceil t_y\rceil}\left(\frac{1}{\pi(i+1)} - \frac{1}{\pi(i)}\right)\frac{1}{\hat{\pi}(-x_i)}.
\end{aligned}$$

Note that because $x_i \geq -x$ for all $i$ in the sum, we have $-x_i \leq x$ and $\frac{1}{\hat{\pi}(-x_i)} \leq \frac{1}{\hat{\pi}(x)}$. Therefore, by telescoping again,

$$\begin{aligned}
\frac{\hat{\zeta}(x)}{\hat{\pi}(x)} + \frac{\hat{\zeta}(y)}{\hat{\pi}(y)} &\leq \frac{1}{\hat{\pi}(x)\hat{\pi}(-x)} + \sum_{i=\lceil t_{-x}\rceil}^{\lceil t_y\rceil}\left(\frac{1}{\pi(i+1)} - \frac{1}{\pi(i)}\right)\frac{1}{\hat{\pi}(x)} \\
&= \frac{1}{\pi(\lceil t_y\rceil)\hat{\pi}(x)} = \frac{1}{\hat{\pi}(y)\hat{\pi}(x)}
\end{aligned}$$

as desired.

Now suppose $y < -x$. Then by telescoping and Claim K.4,

$$\begin{aligned}
\frac{\hat{\zeta}(x)}{\hat{\pi}(x)} + \frac{\hat{\zeta}(y)}{\hat{\pi}(y)} &= \frac{\hat{\zeta}(x)}{\hat{\pi}(x)} + \frac{\hat{\zeta}(-x)}{\hat{\pi}(-x)} - \sum_{i=\lceil t_y\rceil}^{\lceil t_{-x}\rceil}\left(\frac{\zeta_{i+1}}{\pi(i+1)} - \frac{\zeta_i}{\pi(i)}\right) \\
&= \frac{1}{\hat{\pi}(x)\hat{\pi}(-x)} - \sum_{i=\lceil t_y\rceil}^{\lceil t_{-x}\rceil}\left(\frac{1}{\pi(i+1)} - \frac{1}{\pi(i)}\right)\frac{1}{\hat{\pi}(-x_i)}.
\end{aligned}$$

Note that because $x_i \leq -x$ for all $i$ in the sum, we have $-x_i \geq x$ and $\frac{1}{\hat{\pi}(-x_i)} \geq \frac{1}{\hat{\pi}(x)}$. Therefore, by telescoping again,

$$\begin{aligned}
\frac{\hat{\zeta}(x)}{\hat{\pi}(x)} + \frac{\hat{\zeta}(y)}{\hat{\pi}(y)} &\leq \frac{1}{\hat{\pi}(x)\hat{\pi}(-x)} - \sum_{i=\lceil t_y\rceil}^{\lceil t_{-x}\rceil}\left(\frac{1}{\pi(i+1)} - \frac{1}{\pi(i)}\right)\frac{1}{\hat{\pi}(x)} \\
&= \frac{1}{\pi(\lceil t_y\rceil)\hat{\pi}(x)} = \frac{1}{\hat{\pi}(y)\hat{\pi}(x)}
\end{aligned}$$

as desired. ☐

**Claim K.4.** *For any $i = 1, \ldots n-1$, we have*

$$\frac{\zeta_{i+1}}{\pi(i+1)} - \frac{\zeta_i}{\pi(i)} = \frac{1}{\hat{\pi}(-x_i)}\left(\frac{1}{\pi(i+1)} - \frac{1}{\pi(i)}\right).$$

*Proof.* By Eq. (40), for any $x \in \{x_1, \ldots, x_{n-1}\}$, we have

$$\frac{\hat{\zeta}(x+\epsilon)}{\hat{\pi}(x+\epsilon)} - \frac{\hat{\zeta}(x-\epsilon)}{\hat{\pi}(x-\epsilon)} = \frac{1}{\hat{\pi}(x+\epsilon)\hat{\pi}(-x)} - \frac{1}{\hat{\pi}(x-\epsilon)\hat{\pi}(-x)} \geq 0$$

for sufficiently small $\epsilon > 0$. This is because, by Eq. (37), $\hat{\pi}(-x+\epsilon) = \hat{\pi}(-x-\epsilon)$ and $\hat{\zeta}(-x+\epsilon) = \hat{\zeta}(-x-\epsilon)$, and we obtain the above by subtracting Eq. (40) for $x+\epsilon$ and $x-\epsilon$. In particular, letting $i = \lceil t_{x-\epsilon} \rceil$ so that $i+1 = \lceil t_{x+\epsilon} \rceil$, we have the desired claim. □

□

*Remark* K.20. Bobby He provided a much simpler argument to show that $\sum_{i<j} C_{ij}^2 \leq 2r^2 + 6r$ (which is slightly weaker than Remark K.20 but suffices for our downstream needs):

Let $J \stackrel{\text{def}}{=} \{i : \Pi_{ii} \geq 1/2\}$ and $\bar{J} = [n] \setminus J$. Then because $\operatorname{tr} \Pi_{ii} = n - r$, $|\bar{J}| \leq 2r$. We split the squared sum of off-diagonal entries into

$$\sum_{i \neq j} C_{ij}^2 = \sum_{\substack{i \neq j \\ i,j \in J}} C_{ij}^2 + 2\sum_{\substack{i \in J \\ j \in \bar{J}}} C_{ij}^2 + \sum_{\substack{i \neq j \\ i,j \in \bar{J}}} C_{ij}^2.$$

We bound each of the terms separately.

$$\sum_{\substack{i \neq j \\ i,j \in J}} C_{ij}^2 \leq 4 \sum_{\substack{i \neq j \\ i,j \in J}} \Pi_{ij}^2 \leq 4 \sum_{i \neq j} \Pi_{ij}^2 = 4(\operatorname{tr} \Pi^2 - \sum_i \Pi_{ii}^2) \leq 4(k - n(k/n)^2) = 4(n-k)k/n \leq 4r.$$

$$2\sum_{\substack{i \in J \\ j \in \bar{J}}} C_{ij}^2 = 2\sum_{\substack{i \in J \\ j \in \bar{J}}} \Pi_{ij}^2/\Pi_{ii}\Pi_{jj} \leq 4\sum_{\substack{i \in J \\ j \in \bar{J}}} \Pi_{ij}^2/\Pi_{jj} = 4\sum_{j \in \bar{J}} \Pi_{jj}/\Pi_{jj} = 4|\bar{J}| \leq 8r.$$

Above, we used the fact that $\Pi^2 = \Pi \implies \Pi_{jj} + \sum_i \Pi_{ij}^2$. Finally, because $|C_{ij}| \leq 1$, we have

$$\sum_{\substack{i \neq j \\ i,j \in \bar{J}}} C_{ij}^2 \leq |\bar{J}|^2 \leq 4r^2.$$

Summing the above and dividing by two yields the desired result.

### K.5 Law of Large Numbers for Images of Weakly Correlated Gaussians

**Theorem K.21.** *Let $z \sim \mathcal{N}(0, \Pi)$ where $\Pi \in \mathbb{R}^{n \times n}$ is a matrix with nonzero diagonal entries and satisfying $\sum_{i<j} \Pi_{ij}^2/(\Pi_{ii}\Pi_{jj}) \leq R$ for some constant $R$. Consider functions $\phi_i : \mathbb{R} \to \mathbb{R}, i \in [n]$, with finite variance* $\operatorname{Var}(\phi_i(x) : x \sim \mathcal{N}(0, \Pi_{ii})) < \infty$. *Then*

$$\operatorname{Var}\left(\sum_{i=1}^{n} \phi_i(z_i)\right) \leq \left(R\sqrt{2} + 1\right) \sum_i \operatorname{Var}(\phi_i(x) : x \sim \mathcal{N}(0, \Pi_{ii})).$$

Note, by Remark K.20, if $\Pi$ is an orthogonal projection matrix of rank $k$, then we may take $R = 0.5((n-k)^2 + (n-k))$. By Theorem K.21 and Chebyshev's inequality, we have the following.

**Corollary K.22** (Weak Law of Large Numbers for Images of Weakly Correlated Gaussians)**.** *Consider a triangular array $\{\zeta_1^n, \ldots, \zeta_n^n\}_{n \geq 1}$ of Gaussian variables, where each row is given by $\zeta^n \sim \mathcal{N}(0, \Sigma^n)$ and the covariance matrix $\Sigma^n$ has nonzero diagonal entries and satisfies*

$$\sum_{\alpha<\beta} (\Sigma_{\alpha\beta}^n)^2/(\Sigma_{\alpha\alpha}^n \Sigma_{\beta\beta}^n) \leq R(n)$$

*for some $R(n) > 0$. Let $\phi_\alpha^n : \mathbb{R} \to \mathbb{R}, \alpha \in [n], n \geq 1$, be functions with mean $\mu_\alpha^n \stackrel{\text{def}}{=} \mathbb{E}_{\zeta_\alpha^n} \phi_\alpha^n(\zeta_\alpha^n)$. Then, as $n \to \infty$, we have the following convergence in probability*

$$\frac{1}{n} \sum_{\alpha=1}^{n} (\phi_\alpha^n(\zeta_\alpha^n) - \mu_\alpha^n) \xrightarrow{\text{p}} 0$$

*if the following holds*

$$\frac{R(n)\sqrt{2}+1}{n^2}\sum_{\alpha=1}^{n}\operatorname*{Var}_{\zeta_\alpha^n}(\phi_\alpha^n(\zeta_\alpha^n)) \to 0.$$

Note that, if the variance $\mathrm{Var}_{\zeta_\alpha^n}(\phi_\alpha^n(\zeta_\alpha^n))$ is uniformly bounded by some $T > 0$, then $R(n)$ can grow like $n^{1-\varepsilon}$ and we still have convergence in probability.

*Proof of Theorem K.21.* Let $C = D^{-1/2}\Pi D^{-1/2}, D = \mathrm{Diag}(\Pi)$, be the correlation matrix of $\Pi$. The premise of the theorem implies $\sum_{i<j} C_{ij}^2 \le R$. Define functions $\psi_i$ by $\psi_i(y) = \phi_i(\sqrt{\Pi_{ii}}y) - \mathbb{E}_{x\sim\mathcal{N}(0,1)}[\phi_i(\sqrt{\Pi_{ii}}x)]$. Then

$$\operatorname*{Var}_{z\sim\mathcal{N}(0,\Pi)}\left(\sum_{i=1}^n \phi_i(z_i)\right) = \operatorname*{\mathbb{E}}_{z\sim\mathcal{N}(0,C)}\left(\sum_{i=1}^n \psi_i(z_i)\right)^2 = \operatorname*{\mathbb{E}}_{z\sim\mathcal{N}(0,C)}\sum_{i,j}\psi_i(z_i)\psi_j(z_j).$$

Expand $\psi_i$ in the Hermite orthonormal basis,

$$\psi_i(x) = a_{i1}H_1(x) + a_{i2}H_2(x) + \cdots$$

where $H_j(x)$ is the $j$th Hermite polynomial, normalized so that $\mathbb{E}_{z\sim\mathcal{N}(0,1)} H_j(z)^2 = 1$ (note that $H_0(x) = 1$ and does not appear here because $\mathbb{E}_{x\sim\mathcal{N}(0,1)}\psi_i(x) = 0$ by construction). For any locally integrable $\phi : \mathbb{R} \to \mathbb{R}$, let $\|\phi\|_G^2 \overset{\text{def}}{=} \mathbb{E}_{z\sim\mathcal{N}(0,1)}\phi(z)^2$, so that $\|\psi_i\|_G^2 = \sum_k a_{ik}^2 = \mathrm{Var}\left(\phi_i(x) : x \sim \mathcal{N}(0,\Pi_{ii})\right)$. Then,

$$\begin{aligned}
\sum_{i<j}\operatorname*{\mathbb{E}}_{z\sim\mathcal{N}(0,C)}\psi_i(z_i)\psi_j(z_j) &= \sum_{i<j}\sum_{k=1}^{\infty} a_{ik}a_{jk}C_{ij}^k \\
&\le \sum_{k=1}^{\infty}\sqrt{\left(\sum_{i<j} a_{ik}^2 a_{jk}^2\right)\left(\sum_{i<j} C_{ij}^{2k}\right)} \\
&\le \sum_{k=1}^{\infty}\sqrt{\frac{1}{2}\left(\sum_i a_{ik}^2\right)^2\left(\sum_{i<j} C_{ij}^2\right)} && \text{since } |C_{ij}| \le 1 \\
&\le 2^{-1/2}\sum_{k=1}^{\infty}\left(\sum_i a_{ik}^2\right)R && \text{by premise} \\
&= \frac{R}{\sqrt{2}}\sum_i \|\psi_i\|_G^2
\end{aligned}$$

On the other hand, $\sum_i \mathbb{E}_{x\sim\mathcal{N}(0,1)}\psi_i(x)^2 = \sum_i \|\psi\|_G^2$, so that

$$\begin{aligned}
\sum_{i,j}\operatorname*{\mathbb{E}}_{z\sim\mathcal{N}(0,C)}\psi_i(z_i)\psi_j(z_j) &\le \left(R\sqrt{2}+1\right)\sum_i \|\psi_i\|_G^2 \\
&= \left(R\sqrt{2}+1\right)\sum_i \mathrm{Var}\left(\phi_i(x) : x \sim \mathcal{N}(0,\Pi_{ii})\right).
\end{aligned}$$

□

**Theorem K.23.** *Let $z \sim \mathcal{N}(0,\Pi)$ where $\Pi \in \mathbb{R}^{n\times n}$ is a matrix with nonzero diagonal entries and satisfying $\sum_{i<j}\Pi_{ij}^2/(\Pi_{ii}\Pi_{jj}) \le R$ for some constant $R$ independent of $n$. Consider functions $\phi_i : \mathbb{R} \to \mathbb{R}$ for each $i \in [n]$ with mean $\mu_i \overset{def}{=} \mathbb{E}_{z_i}\phi_i(z_i)$. Suppose each $\phi_i$ has finite (2p)th centered moment $\mathbb{E}_{z_i}(\phi_i(z_i) - \mu_i)^{2p}$, where $p \ge 6$. Then for $Q \overset{def}{=} \frac{1}{n}\sum_{i=1}^n \phi_i(z_i)$,*

$$\mathbb{E}[(Q - \mathbb{E}\,Q)^{2p}] \le \mathcal{C}n^{-1.5}\max_{i\in[n]}\mathbb{E}\left(\phi_i(z_i) - \mu_i\right)^{2p}$$

*for some constant $\mathcal{C}$ depending on $p$ and $R$, but not on $n$ or the functions $\phi_i$. If in addition, each $\phi_i$ has finite centered moments of order $2pL$ for some $L > 1$, then*

$$\mathbb{E}[(Q - \mathbb{E}\,Q)^{2p}] \le \mathcal{C} n^{-1.5+1/L} \sqrt[L]{\frac{1}{n}\sum_{i=1}^{n} \mathbb{E}\,(\phi_i(z_i) - \mu_i)^{2pL}}.$$

Note, by Remark K.20, an orthogonal projection matrix of rank $n - O(1)$ satisfies the off-diagonal condition of Theorem K.23. Combining Theorem K.23 and Lemma K.2, we obtain the following.

**Corollary K.24** (Strong Law of Large Numbers for Images of Weakly Correlated Gaussians)**.** *Consider a triangular array $\{\zeta_1^n, \ldots, \zeta_n^n\}_{n\ge 1}$ of Gaussian variables, where each row is given by $\zeta^n \sim \mathcal{N}(0, \Sigma^n)$ and the covariance matrix $\Sigma^n$ has nonzero diagonal entries and satisfies*

$$\sum_{\alpha<\beta} (\Sigma^n_{\alpha\beta})^2 / (\Sigma^n_{\alpha\alpha}\Sigma^n_{\beta\beta}) \le R$$

*for some $R > 0$ independent of $n$. Let $\phi^n_\alpha : \mathbb{R} \to \mathbb{R}, \alpha \in [n], n \ge 1$, be functions with mean $\mu^n_\alpha \overset{\text{def}}{=} \mathbb{E}_{\zeta^n_\alpha} \phi^n_\alpha(\zeta^n_\alpha)$. Then, as $n \to \infty$,*

$$\frac{1}{n}\sum_{\alpha=1}^{n} (\phi^n_\alpha(\zeta^n_\alpha) - \mu^n_\alpha) \xrightarrow{\text{a.s.}} 0$$

*if one of the following condition holds:*

- *For some $p \ge 6$ and some $S$ independent of $n$,*
$$\mathbb{E}_{\zeta^n_\alpha} (\phi^n_\alpha(\zeta^n_\alpha) - \mu^n_\alpha)^{2p} \le S.$$
- *For some $q > 12$ and some $S$ independent of $n$,*
$$\frac{1}{n}\sum_{\alpha=1}^{n} \mathbb{E}_{\zeta^n_\alpha} (\phi^n_\alpha(\zeta^n_\alpha) - \mu^n_\alpha)^{2q} < S.$$

*Proof of Theorem K.23.* Define the correlation matrix $C = D^{-1/2}\Pi D^{-1/2}, D = \mathrm{Diag}(\Pi)$. The premise of the theorem implies $\sum_{i<j} C_{ij}^2 \le R$. Let $\psi_i(y) = \phi_i(\sqrt{\Pi_{ii}}y) - \mathbb{E}_{x\sim\mathcal{N}(0,1)}[\phi_i(\sqrt{\Pi_{ii}}x)]$ be the scaled, centered version of $\phi_i$.

Order the off-diagonal entries of the correlation matrix in the order of decreasing squared value:

$$C^2_{(ij)^{(1)}} \ge C^2_{(ij)^{(2)}} \ge \ldots \ge C^2_{(ij)^{(N)}},$$

where $N = \binom{n}{2}$, and $(ij)^{(t)} = (i^{(t)}j^{(t)})$ are unordered pairs of distinct indices $i^{(t)} \ne j^{(t)}$. Since $\sum_t C^2_{(ij)^{(t)}} \le R$, by Remark K.20, we deduce that

$$|C_{(ij)^{(t)}}| \le n^{-1/4} \qquad \text{for all } t > R\sqrt{n}. \tag{41}$$

Consider the $(2p)$th centered moment

$$\mathbb{E}\left(\frac{1}{n}\sum_{i=1}^{n}\psi_i(y_i)\right)^{2p} = n^{-2p}\,\mathbb{E}\sum_{\sigma:[2p]\to[n]}\prod_{a=1}^{2p}\psi_{\sigma(a)}(y_{\sigma(a)}),$$

where $y \sim \mathcal{N}(0, C)$. We shall bound the sum to show that this moment is not too large.

First note the naive bound via AM-GM,

$$\begin{aligned}\mathbb{E}\left|\prod_{a=1}^{2p}\psi_{\sigma(a)}(y_{\sigma(a)})\right| &\le \mathbb{E}\,\frac{1}{2p}\sum_{a=1}^{2p}\psi_{\sigma(a)}(y_{\sigma(a)})^{2p} \le \max_{i\in[n]} \mathbb{E}_{y\sim\mathcal{N}(0,1)}[\psi_i(y)^{2p}] \\ &= \max_{i\in[n]} \mathbb{E}[(\phi_i(z_i) - \mathbb{E}\,\phi_i(z_i))^{2p} : z_i \sim \mathcal{N}(0, \Pi_{ii})] \overset{\text{def}}{=} B_{2p}.\end{aligned} \tag{42}$$

Now, for any collection of numbers $\{x_i \in \mathbb{R}\}_{i=1}^m$ and any $L > 0$, we have the trivial bound $\max_i |x_i| \leq \left(\sum_{j=1}^m |x_j|^L\right)^{1/L}$, and this bound is tighter the larger $L$ is. Thus

$$B_{2p} \leq n^{1/L} \sqrt[L]{\frac{1}{n}\sum_{i=1}^n (\mathbb{E}[\psi_i(y)^{2p}])^L} \leq n^{1/L} B_{2p,L},$$

where $B_{2p,L} \overset{\text{def}}{=} \sqrt[L]{\frac{1}{n}\sum_{i=1}^n \mathbb{E}[\psi_i(y)^{2pL}]}$, for any $L$.

We can categorize the $n^{2p}$ terms of

$$\sum_{\sigma:[2p]\to[n]} \mathbb{E}\prod_{a=1}^{2p} \psi_{\sigma(a)}(y_{\sigma(a)}) \tag{43}$$

as follows.

- Suppose $\sigma$ is injective.
  - Suppose for each $a \neq b$, $(\sigma(a)\sigma(b)) = (ij)^{(t)}$ for some $t > R\sqrt{n}$, so that $|C_{\sigma(a)\sigma(b)}| \leq n^{-1/4}$ by Eq. (41). By Lemma K.16,

    $$\mathbb{E}\prod_{a=1}^{2p} \psi_{\sigma(a)}(y_{\sigma(a)}) \leq \mathcal{C}_1\left(\prod_{r=1}^{2p}\|\psi_{\sigma(r)}\|_G\right)\left(n^{-1/4}\right)^p$$

    for some constant $\mathcal{C}_1$ dependent on $p$ but *not* on $R$, $\{\psi_r\}_r$, or $n$. Thus the contribution of all such $\sigma$ to the sum Eq. (43) is

    $$\begin{aligned}\sum_{\text{all such }\sigma}\mathcal{C}_1\left(\prod_{r=1}^{2p}\|\psi_{\sigma(r)}\|_G\right)n^{-p/4} &\leq \sum_{\text{all }\sigma}\mathcal{C}_1\left(\prod_{r=1}^{2p}\|\psi_{\sigma(r)}\|_G\right)n^{-p/4}\\ &\leq \mathcal{C}_1 n^{-p/4}\left(\sum_{i=1}^{n}\|\psi_{\sigma(r)}\|_G\right)^{2p}\\ &= \mathcal{C}_1 n^{1.75p} B'_{2p}\end{aligned}$$

    where we have set

    $$B'_{2p} \overset{\text{def}}{=} \left(\frac{1}{n}\sum_{i=1}^n \|\psi_i\|_G\right)^{2p}.$$

  - Suppose for some $a, b \in [2p]$, $(\sigma(a)\sigma(b)) = (ij)^{(t)}$ for $t \leq R\sqrt{n}$. There are at most $2\binom{2p}{2}R\sqrt{n}\cdot n^{2p-2} \leq \mathcal{C}_2 n^{2p-1.5}$ such $\sigma$, for some $\mathcal{C}_2$ depending on $p$ and $R$ but not $n$ (or $\{\psi_r\}_r$). Indeed, there are $R\sqrt{n}$ of choosing such a $t$, $2\binom{2p}{2}$ ways of choosing their preimages under $\sigma$ out of $2p$, and $\leq n^{2p-2}$ ways of choosing the rest of the values of $\sigma$. By Eq. (42), the contribution of all such $\sigma$ to the sum is at most $\mathcal{C}_2 n^{2p-1.5} B_{2p}$.
- Suppose for some $a^* \neq b^*$ in $[2p]$, $\sigma(a^*) = \sigma(b^*)$, but $\sigma|_{[n]\setminus\{a^*,b^*\}}$ is injective and takes range outside $\{\sigma(a^*)\}$. There are $\binom{2p}{2}n\binom{n-1}{2p-2} \leq \mathcal{C}_3 n^{2p-1}$ such $\sigma$, where $\mathcal{C}_3$ depends only on $p$ (but not on $R$, $\{\psi_r\}_r$, and $n$).
  - Suppose for each $a \neq b$, $(\sigma(a)\sigma(b)) = (ij)^{(t)}$ for some $t > R\sqrt{n}$, so that $|C_{\sigma(a)\sigma(b)}| \leq n^{-1/4}$. We apply Lemma K.16 to the $2p-1$ functions $\{\psi^2_{\sigma(a^*)}\} \cup \{\psi_{\sigma(a)}\}_{a\notin\{a^*,b^*\}}$, with $\psi^2_{\sigma(a^*)}$ being the sole function whose expectation is not 0, so that the $I$ of Lemma K.16 has size $2p-2$, and the $\lambda$ of Lemma K.16 is $(2p-2)n^{-1/4}$.

Then Lemma K.16 gives

$$\mathbb{E}\prod_{a=1}^{2p}\psi_{\sigma(a)}(z_{\sigma(a)}) \leq \mathcal{C}_4\|\psi^2_{\sigma(a^*)}\|_G\left(\prod_{a\notin\{a^*,b^*\}}\|\psi_{\sigma(a)}\|_G\right)(n^{-1/4})^{(2p-2)/2}$$

$$= \mathcal{C}_4\|\psi^2_{\sigma(a^*)}\|_G\left(\prod_{a\notin\{a^*,b^*\}}\|\psi_{\sigma(a)}\|_G\right)n^{-(p-1)/4}$$

for some constant $\mathcal{C}_4$ depending on $p$ but *not* on $n$, $R$, or $\{\psi_r\}_r$. Thus the collective contribution of such $\sigma$ to the sum is at most

$$\sum_{\text{all such }\sigma}\mathcal{C}_4\|\psi^2_{\sigma(a^*)}\|_G\left(\prod_{a\notin\{a^*,b^*\}}\|\psi_{\sigma(a)}\|_G\right)n^{-(p-1)/4}$$

$$\leq \mathcal{C}_4 n^{-(p-1)/4}\binom{p}{2}\sum_{i=1}^{n}\|\psi_i^2\|_G\sum_{\pi:[2p-2]\to[n]\setminus\{i\}}\prod_{a=1}^{2p-2}\|\psi_{\pi(a)}\|_G$$

$$\leq \mathcal{C}_4 n^{-(p-1)/4}\binom{p}{2}\sum_{i=1}^{n}\|\psi_i^2\|_G\sum_{\pi:[2p-2]\to[n]}\prod_{a=1}^{2p-2}\|\psi_{\pi(a)}\|_G$$

$$= \mathcal{C}_5 n^{-(p-1)/4}\left(\sum_{i=1}^{n}\|\psi_i^2\|_G\right)\left(\sum_{i=1}^{n}\|\psi_i\|_G\right)^{2p-2}$$

$$= \mathcal{C}_5 n^{1.75(p-1)}B''_{2p},$$

where $\mathcal{C}_5 = \mathcal{C}_4\binom{p}{2}$ and depends only on $p$, but *not* on $n$, $R$, or $\{\psi_r\}_r$, and where we have set

$$B''_{2p} = \left(\frac{1}{n}\sum_{i=1}^{n}\|\psi_i^2\|_G\right)\left(\frac{1}{n}\sum_{i=1}^{n}\|\psi_i\|_G\right)^{2p-2}.$$

  – Suppose for some $a, b \in [2p]$, $(\sigma(a)\sigma(b)) = (ij)^{(t)}$ for $t \leq R\sqrt{n}$. There are at most $\binom{2p}{2}n\cdot R\sqrt{n}\cdot\binom{n-2}{2p-3} = \mathcal{C}_6 n^{2p-1.5}$ such $\sigma$, where $\mathcal{C}_6$ depends only on $p$ and $R$, but *not* on $n$ or $\{\psi_r\}_r$. Using Eq. (42) again, we can upper bound the contribution of such $\sigma$ by $\mathcal{C}_6 n^{2p-1.5}B_{2p}$.

- Otherwise, there are more than one pair of inputs that collide under $\sigma$. There are at most $\mathcal{C}_7 n^{2p-2}$ such $\sigma$, where $\mathcal{C}_7$ depends only on $p$, but *not* on $n$, $R$, or $\{\psi_r\}_r$. Using Eq. (42), we upper bound their contributions by $\mathcal{C}_7 n^{2p-2}B_{2p}$.

To summarize, using $O(-)$ to hide the constants $\mathcal{C}_*$ which don't depend on $n$ or the functions $\psi_i$:

$$\mathbb{E}\sum_{\sigma:[2p]\to[n]}\prod_{a=1}^{2p}\psi_{\sigma(a)}(y_{\sigma(a)}) \leq O\left(n^{1.75p}B'_{2p} + n^{1.75(p-1)}B''_{2p} + n^{2p-1.5}B_{2p}\right)$$

$$\leq O\left(n^{1.75p}B'_{2p} + n^{1.75(p-1)}B''_{2p} + n^{2p-1.5+1/L}B_{2p,L}\right)$$

$$\mathbb{E}\left(\frac{1}{n}\sum_{i=1}^{n}\phi_i(z_i)-\mu_i\right)^{2p} \leq O\left(n^{-0.25p}B'_{2p} + n^{-0.25p-1.75}B''_{2p} + n^{-1.5}B_{2p}\right)$$

$$\leq O\left(n^{-0.25p}B'_{2p} + n^{-0.25p-1.75}B''_{2p} + n^{-1.5+1/L}B_{2p,L}\right)$$

By the power mean inequality, we get that $B'_{2p}, B''_{2p} \leq B_{2p} \leq n^{1/L}B_{2p,L}$. Substitution then gives the desired result.

□

This theorem can be significantly strengthened with more careful case work and applying more involved versions of Lemmas K.16 and K.17, but we will not be concerned with this here.

## L Proof of NETSOR$\top$ Master Theorem

In this section, we will give the proof for our main theorem.

**A Bit of Notation and Terminology** Note that, for each $n$, the randomness of our program specified by Theorem 2.10 comes from the sampling of the initial vectors $\mathcal{V}$ and matrices $\mathcal{W}$. Let $\mathsf{U}$ be the product space obtained from multiplying together the corresponding probability space for each $n$. Each sample from this product probability space thus correspond to a sequence $\{S(n)\}_n$ of instantiatiations of $\mathcal{V}$ and $\mathcal{W}$. Below, when we say "almost surely" (often abbreviated "a.s."), we mean "almost surely over the probability of $\mathsf{U}$." We will also often make statements of the form

$$\textit{almost surely (or, a.s.), for all large } n, \quad \mathcal{A}(n) \textit{ is true}$$

where $\mathcal{A}(n)$ is a claim parametrized by $n$. This means that for all but a $\mathsf{U}$-probability-zero set of sequences $\{S(n)\}_n$, $\mathcal{A}(n)$ is true for large enough $n$. Note that the order of the qualifiers is very important here.

Let $g^1, \dots, g^M$ be all of the G-vars (including those of $\mathcal{V}$) in the program. It suffices to prove Theorem 2.10 where $k = M$ and the vector $\{h^i\}_i$ are the G-vars $g^1, \dots, g^M$, since all other vectors are `Nonlin` images of them. For convenience, we also assume that for all $g \in \mathcal{V}$, we have $\mathbb{E}\, Z^g = 0$; this is without loss of generality (WLOG) since any nonzero mean can be absorbed into applications of `Nonlin`.

**We induct, but on what?** A natural way of going about proving Theorem 2.10 is by inducting on the number of variables in a program. While this would be the main backbone of the proof, it turns out such a proof would require us to assume a certain rank condition (Lemma L.11) on $g^1, \dots, g^M$ which is not so easy to check. Thus, it would be more fruitful to perform a simultaneous induction on our claim (Moments) along with another statement, parametrized by $m$, that would prove such a condition for us.

**Moments**$(m)$ For any polynomially-bounded $\psi : \mathbb{R}^m \to \mathbb{R}$, as $n \to \infty$,

$$\frac{1}{n} \sum_{\alpha=1}^{n} \psi(g^1_\alpha, \dots, g^m_\alpha) \xrightarrow{\text{a.s.}} \mathbb{E}\, \psi\left(Z^{g^1}, \dots, Z^{g^m}\right).$$

**CoreSet**$(m)$ There exists a "core set" $\mathcal{M} \subseteq [m]$ such that,

**Basis**$(m)$ almost surely, for large enough $n$, for every $i \in [m]$, there exist *unique* constants (not depending on $n$) $\{a_j\}_{j \in \mathcal{M}}$ such that $g^i = \sum_{j \in \mathcal{M}} a_j g^j$. Note the uniqueness implies that $\{g^i\}_{i \in \mathcal{M}}$ is linearly independent.

**Density**$(m)$ The distribution of the random vector

$$\{Z^{g^j}\}_{j \in \mathcal{M}} \in \mathbb{R}^{\mathcal{M}}$$

is absolutely continuous w.r.t. the Lebesgue measure in $\mathbb{R}^{\mathcal{M}}$, and vice versa.[45] In other words, for every measurable set $U \subseteq \mathbb{R}^{\mathcal{M}}$,

$$\Pr[\{Z^{g^j}\}_{j \in \mathcal{M}} \in U] = 0 \iff U \text{ has Lebesgue measure } 0.$$

**NullAvoid**$(m)$ for every triangular array of Lesbegue measure zero sets $\{U_{n\alpha} \in \mathbb{R}^{\mathcal{M}}\}_{n \in \mathbb{N}, \alpha \in [n]}$, almost surely for all large enough $n$, for all $\alpha \in [n]$, we have

$$\{g^i_\alpha\}_{i \in \mathcal{M}} \not\in U_{n\alpha}.$$

In other words, the values $\{g^i_\alpha\}_{\alpha \in \mathcal{M}}$ of the core set "avoid" Lebesgue measure zero sets asymptotically. Intuitively, this says that the distribution of these values are not singular. (Note the LHS depends on $n$ although we are suppressing it notationally)

Let us explain in brief why we need to consider CoreSet satisfying Basis and NullAvoid.

[45] Compared to the proof of the Master Theorem in Yang [51], this is a new inductive hypothesis. This is trivially true in the NTK case since the distribution in question is Gaussian with full support.

- Basis reduces the consideration of Moments to only the core set G-vars, since every other G-var is asymptotically a linear combination of them.
- When we apply the Gaussian conditioning technique Proposition K.3, we need to reason about the pseudo-inverse $\Lambda^+$ of some covariance matrix $\Lambda$. Each entry of $\Lambda$ is of the form $\frac{1}{n}\sum_{\alpha=1}^{n}\phi_i(g^1_\alpha,\ldots,g^{m-1}_\alpha)\phi_j(g^1_\alpha,\ldots,g^{m-1}_\alpha)$ for a collection of polynomially bounded scalar functions $\{\phi_i\}_i$. This $\Lambda$ will be a random variable which converges a.s. to a determinstic limit $\mathring{\Lambda}$ as $n\to\infty$. It should be generically true that $\Lambda^+ \xrightarrow{\text{a.s.}} \mathring{\Lambda}^+$ as well, which is essential to make the Gaussian conditioning argument go through. But in general, this is guaranteed only if $\Lambda$'s rank doesn't drop suddenly in the $n\to\infty$ limit. We thus need to guard against the possibility that $g^1,\ldots,g^m$, in the limit, suddenly concentrate on a small set on which $\{\phi_i(g^1,\ldots,g^m)\}_i$ are linearly dependent. This is where NullAvoid comes in. It tells us that $g^1,\ldots,g^m$ will avoid any such small set asymptotically, so that indeed the rank of $\Lambda$ will not drop in the limit.

**Proof organization** We will show that Moments and CoreSet are true for initial G-vars in $\mathcal{V}$, as the base case, and

$$\text{Moments}(m-1) \text{ and } \text{CoreSet}(m-1) \implies \text{Moments}(m) \text{ and } \text{CoreSet}(m)$$

as the inductive step. By induction, we obtain Moments$(M)$, which is Theorem 2.10.

The base cases are easy and we will dispatch with them immediately after this in Appendix L.3, but the inductive step is much more complicated, and we will need to set up notation in Appendix L.4. During this setup, we prove some basic limit theorems using the induction hypothesis. However, the full generality of these claims requires some consequences of CoreSet, which we call "rank stability" and "zero stability." These notions are introduced and proved in Appendix L.5.

We would then finally be able to handle the inductive steps at this point. We first prove

$$\text{Moments}(m-1) \text{ and } \text{CoreSet}(m-1) \implies \text{CoreSet}(m)$$

in Appendix L.6 because it is easier. Then we prove

$$\text{Moments}(m-1) \text{ and } \text{CoreSet}(m-1) \implies \text{Moments}(m)$$

in Appendix L.7.

Before we proceed with the induction, let us set up establish some matrix notations that will make the proof significantly easier to express.

### L.1 Preliminaries: Square-Integrable Random Variables and Matrix Notation

**Definition L.1.** Consider the space $\mathcal{L}$ of square-integrable random variables. This space has inner product given by $(X,Y)\mapsto \mathbb{E}\,XY$ for any $X,Y\in\mathcal{L}$. We will often use matrix notation in *square brackets* to express certain sum of (inner) products between elements of $\mathcal{L}$ and real scalars in $\mathbb{R}$. For example, for $X^1,\ldots,X^k,Y^1,\ldots,Y^k,X,Y\in\mathcal{L}$ and $a_1,\ldots,a_k,a\in\mathbb{R}$, we write

$$[X^1,\ldots,X^k]\begin{bmatrix}Y^1\\ \vdots\\ Y^k\end{bmatrix} = \sum_{i=1}^{k}\mathbb{E}\,X^iY^i\in\mathbb{R} \qquad [Y^1,\ldots,Y^k]\begin{bmatrix}a_1\\ \vdots\\ a_k\end{bmatrix} = \sum_{i=1}^{k}a_iY^i\in\mathcal{L}$$

$$\begin{bmatrix}X^1\\ \vdots\\ X^k\end{bmatrix}[Y] = \begin{bmatrix}\mathbb{E}\,X^1Y\\ \vdots\\ \mathbb{E}\,X^kY\end{bmatrix}\in\mathbb{R}^k \qquad \begin{bmatrix}X^1\\ \vdots\\ X^k\end{bmatrix}[a] = \begin{bmatrix}aX^1\\ \vdots\\ aX^k\end{bmatrix}\in\mathcal{L}^k.$$

In addition,

$$\begin{bmatrix}X^1\\ \vdots\\ X^k\end{bmatrix}[Y^1,\ldots,Y^k] = \begin{bmatrix}\mathbb{E}\,X^1Y^1 & \cdots & \mathbb{E}\,X^1Y^k\\ \vdots & \ddots & \vdots\\ \mathbb{E}\,X^kY^1 & \cdots & \mathbb{E}\,X^kY^k\end{bmatrix}\in\mathbb{R}^{k\times k},$$

but note that this "outer product" is not a rank-1 matrix, but rather full rank typically. In general, a matrix in $\mathcal{L}^{k\times l}$ multiplied by a matrix $\mathcal{L}^{l\times m}$ on the right produces a matrix $\mathbb{R}^{k\times m}$, where the elementwise product is provided by the inner product of $\mathcal{L}$; a matrix in $\mathcal{L}^{k\times l}$ multiplied by a matrix $\mathbb{R}^{l\times m}$ on the right produces a matrix $\mathcal{L}^{k\times m}$, where the elementwise product is provided by the scalar multiplication of $\mathcal{L}$. In general, this mixed-type matrix multiplication is not associative, i.e. $A(BC)\neq(AB)C$. We will always read a series of matrix multiplication from the right: $ABCD=A(B(CD))$.

**Fact L.2.** *Just like in a finite-dimensional inner product space, given a finite collection $\mathcal{S}=\{X^i\}_{i=1}^k\subseteq\mathcal{L}$, the orthogonal projection operator $\Pi_{\mathcal{S}}$ to the span of $\mathcal{S}$ (inside $\mathcal{L}$) is given by*

$$\Pi_{\mathcal{S}}Y\overset{def}{=}\sum_{i=1}^k a_iX^i,$$

*for any $Y\in\mathcal{L}$, where*

$$a=\Lambda^+b\in\mathbb{R}^k,\quad b_j=\mathbb{E}\,X^jY,\ b\in\mathbb{R}^k,\quad \Lambda_{ij}=\mathbb{E}\,X^iX^j,\ \Lambda\in\mathbb{R}^{k\times k}.$$

*In the matrix notation of Definition L.1, this can be expressed as*

$$\begin{aligned}\Pi_{\mathcal{S}}Y&=[X^1,\dots,X^k]\begin{bmatrix}&&\\&\Lambda^+&\\&&\end{bmatrix}\begin{bmatrix}X^1\\ \vdots\\ X^k\end{bmatrix}[Y]\\&=[X^1,\dots,X^k]\left(\begin{bmatrix}X^1\\ \vdots\\ X^k\end{bmatrix}[X^1,\dots,X^k]\right)^+\begin{bmatrix}X^1\\ \vdots\\ X^k\end{bmatrix}[Y].\end{aligned}\tag{44}$$

*We denote the the projection to the orthogonal complement of the linear span of $\mathcal{S}$ by*

$$\Pi_{\mathcal{S}}^{\perp}\overset{def}{=}I-\Pi_{\mathcal{S}}.$$

### L.2 Preliminaries: Adjunction Relation Between $\hat{Z}$ and $\dot{Z}$

A sort of "adjunction" holds between $\hat{Z}$ and $\dot{Z}$, as we describe below. This is a crucial calculation needed to prove that the "Onsager correction term" $\dot{Z}$ is right.

**Lemma L.3.** *For any vectors $x,h\in\mathbb{R}^n$ and matrix $W\in\mathbb{R}^{n\times n}$ in the program, we have*[46]

$$\mathbb{E}\,\hat{Z}^{W^\top x}Z^h=\mathbb{E}\,Z^x\dot{Z}^{Wh}.$$

Compare with the identity

$$\mathbb{E}\,Z^{W^\top x}Z^h=\mathbb{E}\,Z^xZ^{Wh}$$

which, assuming Theorem 2.10, would follow from $(W^\top x)^\top h=x^\top(Wh)$.

*Proof.* Fix $g=Wh$ and we will prove, for all $u=W^\top v$ in the program,

$$\mathbb{E}\,\hat{Z}^uZ^h=\mathbb{E}\,Z^v\dot{Z}^g.$$

Let $\mathcal{G}$ be the set of G-vars introduced before $g$ (so $g\notin\mathcal{G}$). Note that by Proposition I.7, $Z^h$ is a deterministic function of $\hat{Z}^{\mathcal{G}}\overset{\text{def}}{=}\{\hat{Z}^x:x\in\mathcal{G}\}$. Let $\mathcal{E}$ be a subset of $\mathcal{G}$ such that $\hat{Z}^{\mathcal{E}}\overset{\text{def}}{=}\{\hat{Z}^x:x\in\mathcal{E}\}$ has a nonsingular covariance matrix and generates the same $\sigma$-algebra as $\hat{Z}^{\mathcal{G}}$. Then, because $\hat{Z}^{\mathcal{G}}$ is jointly Gaussian, $\hat{Z}^{\mathcal{G}}$ is a linear function of $\hat{Z}^{\mathcal{E}}$, and $Z^h$ is a deterministic function of $\hat{Z}^{\mathcal{E}}$.

Suppose $u^1,\dots,u^k$ are all G-vars in $\mathcal{G}$ introduced via `MatMul` with $W^\top$

$$u^i=W^\top v^i$$

for H-vars $v^1,\dots,v^k$. Assume WLOG that for $\ell\le k$, we have $\{u^1,\dots,u^\ell\}=\{u^1,\dots,u^k\}\cap\mathcal{E}$, so that $\hat{Z}^{u^1},\dots,\hat{Z}^{u^\ell}$ also have full rank covariance. Note that any $\hat{Z}^{u^j},j\in[k]$, has to be a linear combination of $\{\hat{Z}^{u^i}\}_{i=1}^{\ell}$.

[46] In the variable dimension case, if $x\in\mathbb{R}^m,h\in\mathbb{R}^n,W\in\mathbb{R}^{m\times n}$, and $n/m\to\rho$, then we have $\rho\,\mathbb{E}\,\hat{Z}^{W^\top x}Z^h=\mathbb{E}\,Z^x\dot{Z}^{Wh}$.

Any G-var $\bar{g}$ where $\bar{g} = W^\top \bar{h}$ for some $\bar{h}$ must be one of $u^i, i \in [k]$, or $\bar{g} \notin \mathcal{E}$. We will divide up the proof of Lemma L.3 into two cases.

*Case 1.* For any $i \in [k]$, we have $\mathbb{E}\, \hat{Z}^{u^i} Z^h = \mathbb{E}\, Z^{v^i} \dot{Z}^g$.

*Case 2.* For any G-var $\bar{g} \notin \mathcal{E}$ such that $\bar{g} = W^\top \bar{h}$ for some $\bar{h}$, we have $\mathbb{E}\, \hat{Z}^{\bar{g}} Z^h = \mathbb{E}\, Z^{\bar{h}} \dot{Z}^g$.

*Proof of Case 1.* By treating $Z^h$ as a deterministic function of $\hat{Z}^{\mathcal{E}}$, Lemma K.5 says there are coefficients $\{a_x \in \mathbb{R} : x \in \mathcal{E}\}$ such that

$$\mathbb{E}\, \hat{Z}^{u^i} Z^h = \sum_{x \in \mathcal{E}} a_x \operatorname{Cov}(\hat{Z}^{u^i}, \hat{Z}^x).$$

However, if $x \in \mathcal{E} \setminus \{u^1, \ldots, u^k\}$ (i.e. $x \neq W^\top y$ for some $y$), then $\operatorname{Cov}(\hat{Z}^{u^i}, \hat{Z}^x) = 0$ by `ZHat` for all $i = 1, \ldots, k$. Thus the sum above over $x \in \mathcal{E}$ reduces to a sum over $x \in \mathcal{E} \cap \{u^1, \ldots, u^k\} = \{u^1, \ldots, u^\ell\}$:

$$\mathbb{E}\, \hat{Z}^{u^i} Z^h = \sum_{j=1}^{\ell} a_{u^j} \operatorname{Cov}(\hat{Z}^{u^i}, \hat{Z}^{u^j}) = \sum_{j=1}^{\ell} a_{u^j} \sigma_W^2\, \mathbb{E}\, Z^{v^i} Z^{v^j} = \mathbb{E}\, Z^{v^i} \sum_{j=1}^{\ell} a_{u^j} \sigma_W^2 Z^{v^j}.$$

Note that this argument works for any $u^i$ with $i \leq \ell$, and the constants $\{a_{u^j}\}_{j=1}^{\ell}$ *remain the same* for all $u^i$:

$$\forall i \in [\ell], \quad \mathbb{E}\, \hat{Z}^{u^i} Z^h = \mathbb{E}\, Z^{v^i} \sum_{j=1}^{\ell} a_{u^j} \sigma_W^2 Z^{v^j}.$$

By the observation above that any $\hat{Z}^{u^j}, j \in [k]$, is a linear combination of $\{\hat{Z}^{u^i}\}_{i=1}^{\ell}$, this equality is also extended to all $i \in [k]$:

$$\forall i \in [k], \quad \mathbb{E}\, \hat{Z}^{u^i} Z^h = \mathbb{E}\, Z^{v^i} \sum_{j=1}^{\ell} a_{u^j} \sigma_W^2 Z^{v^j}. \tag{45}$$

Set $Y \stackrel{\text{def}}{=} \sum_{j=1}^{\ell} a_{u^j} \sigma_W^2 Z^{v^j}$. To prove our claim for case 1, it suffices to show that $\dot{Z}^g = Y$.

Recall that in Remark 2.12, the matrix $C \in \mathbb{R}^{k \times k}$ and $b \in \mathbb{R}^k$, in the matrix notation of Definition L.1, are given by

$$C = \begin{bmatrix} Z^{v^1} \\ \vdots \\ Z^{v^k} \end{bmatrix} [Z^{v^1}, \ldots, Z^{v^k}], \quad b = \begin{bmatrix} \hat{Z}^{u^1} \\ \vdots \\ \hat{Z}^{u^k} \end{bmatrix} [Z^h].$$

By Eq. (45), we can also re-express $b$ as the following product of a $k \times 1$ vector and a $1 \times 1$ vector,

$$b = \begin{bmatrix} Z^{v^1} \\ \vdots \\ Z^{v^k} \end{bmatrix} [Y].$$

Thus, from Remark 2.12,

$$\dot{Z}^g = [Z^{v^1}, \ldots, Z^{v^k}] \begin{bmatrix} \quad C^+ \quad \end{bmatrix} \begin{bmatrix} b \end{bmatrix} = [Z^{v^1}, \ldots, Z^{v^k}] \begin{bmatrix} \quad C^+ \quad \end{bmatrix} \begin{bmatrix} Z^{v^1} \\ \vdots \\ Z^{v^k} \end{bmatrix} [Y] = \Pi Y$$

where $\Pi$ is the orthogonal projection in $\mathcal{L}$ to the subspace spanned by $\{Z^{v^1}, \ldots, Z^{v^k}\}$, as discussed in Fact L.2. Since $Y$ is already in this subspace, we just get

$$\dot{Z}^g = Y$$

as desired.

*Proof of Case 2.* Let $U$ denote the random column vector $(Z^{u^1}, \dots, Z^{u^k})^\top$. Then, conditioned on $U$, the random variable $Z^h$ has randomness coming only from $\{\hat{Z}^x : x \in \mathcal{E} \setminus \{u^1, \dots, u^k\}\}$. So conditioned on $U$, $Z^h$ is independent from $\hat{Z}^{\bar{g}}$, by `ZHat`. Therefore,

$$\mathbb{E}\, \hat{Z}^{\bar{g}} Z^h = \mathop{\mathbb{E}}_U \left( \mathbb{E}[Z^h \mid U] \, \mathbb{E}[\hat{Z}^{\bar{g}} \mid U] \right). \tag{46}$$

We now will massage the RHS into the desired form. Let $C \in \mathbb{R}^{k\times k}, C_{ij} = \mathrm{Cov}(\hat{Z}^{u^i}, \hat{Z}^{u^j})$ be the covariance matrix of $\hat{Z}^{u^i}$, and let $b \in \mathbb{R}^k, b_i = \mathrm{Cov}(\hat{Z}^{\bar{g}}, \hat{Z}^{u^i})$. Then by standard Gaussian conditioning formula (Proposition K.3), we have

$$\mathbb{E}[\hat{Z}^{\bar{g}} \mid U] = b^\top C^+ U,$$

as a linear function of $U$. Plugging into Eq. (46), in the matrix notation of Definition L.1, we have

$$\mathbb{E}\, \hat{Z}^{\bar{g}} Z^h = b^\top C^+ \begin{bmatrix} \hat{Z}^{u^1} \\ \vdots \\ \hat{Z}^{u^k} \end{bmatrix} [Z^h].$$

By Case 1, we have $\mathbb{E}\, \hat{Z}^{u^i} Z^h = \mathbb{E}\, Z^{v^i} \dot{Z}^g$, so we can write

$$\mathbb{E}\, \hat{Z}^{\bar{g}} Z^h = b^\top C^+ \begin{bmatrix} \hat{Z}^{v^1} \\ \vdots \\ \hat{Z}^{v^k} \end{bmatrix} [\dot{Z}^g] = \mathbb{E} \left( b^\top C^+ V \right) \dot{Z}^g,$$

where $V$ is the random column vector $(Z^{v^1}, \dots, Z^{v^k})^\top$. Now, because $C_{ij} = \sigma_W^2\, \mathbb{E}\, Z^{v^i} Z^{v^j}$ and $b_i = \sigma_W^2\, \mathbb{E}\, Z^{\bar{h}} Z^{v^i}$, we see

$$b^\top C^+ V = \Pi Z^{\bar{h}},$$

where $\Pi$ is the orthogonal projection to the subspace of $\mathcal{L}$ spanned by $\{Z^{v^1}, \dots, Z^{v^k}\}$. Putting this all together, we have, by the self-adjoint property of $\Pi$,

$$\mathbb{E}\, \hat{Z}^{\bar{g}} Z^h = \mathbb{E}(\Pi Z^{\bar{h}}) \dot{Z}^g = \mathbb{E}\, Z^{\bar{h}} (\Pi \dot{Z}^g).$$

However, since $\dot{Z}^g$ is already a linear combination of $\{Z^{v^1}, \dots, Z^{v^k}\}$ by definition (`ZDot`), we have $\Pi \dot{Z}^g = \dot{Z}^g$. Thus we obtain

$$\mathbb{E}\, \hat{Z}^{\bar{g}} Z^h = \mathbb{E}\, Z^{\bar{h}} \dot{Z}^g,$$

as desired. ☐

**Lemma L.4.** *Consider a G-var $g = Wh$ in our program. Suppose $g^i = W^\top h^i, i = 1, \dots, \ell$, includes all G-vars of the form $W^\top x$ introduced before $h$. If we set $Z^H$ be the column vector $[Z^{h^1}, \dots, Z^{h^\ell}]^\top$ and define $C \in \mathbb{R}^{\ell\times\ell}, b \in \mathbb{R}^\ell$ by*

$$C_{ij} \overset{\text{def}}{=} \mathbb{E}\, Z^{h^i} Z^{h^j}, \quad b_i \overset{\text{def}}{=} \mathbb{E}\, \hat{Z}^{g^i} Z^h,$$

*then, using the matrix notation of Definition L.1, we have*

$$\dot{Z}^g = Z^{H\top} C^+ b = [Z^{h^1}, \dots, Z^{h^\ell}] \begin{bmatrix} & & \\ & C^+ & \\ & & \end{bmatrix} \begin{bmatrix} \\ b \\ \\ \end{bmatrix}. \tag{47}$$

Note, by assumption, $\ell$ here is at least $k$ in Remark 2.12. If $\{g^i\}_i$ are exactly the G-vars of the form $W^\top x$ introduced before $h$, then Eq. (47) is equivalent to Eq. (4) in Remark 2.12. So this lemma says this expression always evaluates to $\dot{Z}^g$ as long as $\{g^i\}_i$ is no smaller.

*Proof.* By Lemma L.3, we can re-express

$$b_i = \mathbb{E}\, Z^{h^i} \dot{Z}^g.$$

Then by Fact L.2,

$$Z^{H\top} C^+ b = Z^{H\top} C^+\, \mathbb{E}\, Z^H \dot{Z}^g = \Pi_H \dot{Z}^g$$

where $\Pi_H$ is the orthogonal projection operator to the span of $\{Z^{h^1}, \dots, Z^{h^\ell}\}$; see Fact L.2. But by `ZDot`, $\dot{Z}^g$ is already a linear combination of $\{Z^{h^i} : g^i \text{ introduced before } h\}$. Thus

$$\Pi_H \dot{Z}^g = \dot{Z}^g$$

as desired. ☐

### L.3 Base Cases: Moments and CoreSet for Initial Vectors

**Base case: Moments($\mathcal{V}$)** Since the vectors in $\mathcal{V}$ are sampled iid coordinatewise, Moments($\mathcal{V}$) just follows from the strong law of large numbers.

**Base Case: CoreSet($\mathcal{V}$)** Pick the core set $\mathcal{M}$ to be (the indices of) any subset $\mathcal{U} \subseteq \mathcal{V}$ such that $Z^{\mathcal{U}}$ forms a linear basis of $Z^{\mathcal{V}}$. Then it's straightforward to verify Basis, Density, and NullAvoid.

### L.4 Inductive Case: Setup

We now assume Moments$(m-1)$ and CoreSet$(m-1)$ and want to reason about $g^m$ to show Moments$(m)$ and CoreSet$(m)$. In this section, we will set up the notation and helper lemmas toward this goal. We need to (unfortunately) introduce a large number of symbols. To alleviate possible confusion, we provide an index of all such symbols in Appendix L.4.3.

#### L.4.1 Definitions

Suppose

$$g^m = Ah \quad \text{and} \quad h = \phi(g^1, \ldots, g^{m-1}) \tag{48}$$

for some polynomially-bounded $\phi$. For brevity, we will just write $g = g^m$. Consider all previous instances where $A$ or $A^\top$ is used:

$$x^i = Ay^i, i = 1, \ldots, r, \quad \text{and} \quad u^j = A^\top v^j, j = 1, \ldots, s. \tag{49}$$

Define the matrices $\mathbf{X} \in \mathbb{R}^{n\times r}, \mathbf{U} \in \mathbb{R}^{n\times s}, \mathbf{Y} \in \mathbb{R}^{n\times r}, \mathbf{V} \in \mathbb{R}^{n\times s}$ with $x^i, u^i, y^i, v^i$ as columns

$$\mathbf{X} \stackrel{\text{def}}{=} [x^1|\ldots|x^r], \qquad \mathbf{U} \stackrel{\text{def}}{=} [u^1|\ldots|u^s], \qquad \mathbf{Y} \stackrel{\text{def}}{=} [y^1|\ldots|y^r], \qquad \mathbf{V} \stackrel{\text{def}}{=} [v^1|\ldots|v^s]. \tag{50}$$

Let $\mathcal{B}$ be the $\sigma$-algebra spanned by all previous G-vars $g^1, \ldots, g^{m-1}$; since all previous vectors are deterministic images of these G-vars, $\mathcal{B}$ is also the $\sigma$-algebra generated by all previous vectors before $g$. Conditioning on $\mathcal{B}$, $A$ is linearly constrained by $\mathbf{X} = A\mathbf{Y}, \mathbf{U} = A^\top\mathbf{V}$. Thus we have, by the Gaussian conditioning trick (Lemma K.9),

$$g \stackrel{\text{d}}{=}_{\mathcal{B}} (E + \mathbf{\Pi}_{\mathbf{V}}^{\perp}\tilde{A}\mathbf{\Pi}_{\mathbf{Y}}^{\perp})h \tag{51}$$

where $\tilde{A}$ is an independent copy of $A$, and $\mathbf{\Pi}_{\mathbf{Y}} \stackrel{\text{def}}{=} \mathbf{Y}\mathbf{Y}^+ = \mathbf{Y}(\mathbf{Y}^\top\mathbf{Y})^+\mathbf{Y}^\top$ is the projection to the column space of $\mathbf{Y}$ (likewise for $\mathbf{\Pi}_{\mathbf{V}}$), and

$$\begin{aligned} E &\stackrel{\text{def}}{=} \mathbf{X}\mathbf{Y}^+ + \mathbf{V}^{+\top}\mathbf{U}^\top - \mathbf{V}^{+\top}\mathbf{U}^\top\mathbf{Y}\mathbf{Y}^+ \\ &= \mathbf{X}(\mathbf{Y}^\top\mathbf{Y})^+\mathbf{Y}^\top + \mathbf{V}(\mathbf{V}^\top\mathbf{V})^+\mathbf{U}^\top - \mathbf{V}(\mathbf{V}^\top\mathbf{V})^+\mathbf{U}^\top\mathbf{Y}(\mathbf{Y}^\top\mathbf{Y})^+\mathbf{Y}^\top. \end{aligned} \tag{52}$$

We can make obvious the conditional distribution of $g$ if we rewrite Eq. (51) as

$$g \stackrel{\text{d}}{=}_{\mathcal{B}} \omega + \sigma\mathbf{\Pi}_{\mathbf{V}}^{\perp}z, \quad \text{with } z \sim \mathcal{N}(0, I_n) \tag{53}$$

$$\text{with} \quad \omega \stackrel{\text{def}}{=} Eh \in \mathbb{R}^n, \quad \sigma \stackrel{\text{def}}{=} \sigma_A\sqrt{\|\mathbf{\Pi}_{\mathbf{Y}}^{\perp}h\|^2/n} \in \mathbb{R}. \tag{54}$$

where $\omega, \sigma, \mathbf{\Pi}_{\mathbf{V}}^{\perp}$ are all deterministic conditioned on $\mathcal{B}$, and the only randomness after conditioning comes from $z$. We will see below that $\omega$ can be expressed explicitly as a linear combination in $\mathbf{X}$ and $\mathbf{V}$ (Lemma L.7), and $\sigma$ converges to a deterministic limit $\mathring{\sigma}$ (Lemma L.6). But to get there, we need to first define a few useful matrices and vectors of fixed dimensions

$$\begin{aligned} &\Upsilon \stackrel{\text{def}}{=} \mathbf{Y}^\top\mathbf{Y}/n \in \mathbb{R}^{r\times r} \quad \Lambda \stackrel{\text{def}}{=} \mathbf{V}^\top\mathbf{V}/n \in \mathbb{R}^{s\times s} \quad \Gamma \stackrel{\text{def}}{=} \mathbf{U}^\top\mathbf{Y}/n \in \mathbb{R}^{s\times r} \\ &\gamma \stackrel{\text{def}}{=} \mathbf{Y}^\top h/n \in \mathbb{R}^r \qquad \delta \stackrel{\text{def}}{=} \mathbf{U}^\top h/n \in \mathbb{R}^s. \end{aligned} \tag{55}$$

By induction hypothesis Moments$(m-1)$, $\Upsilon, \Lambda, \Gamma, \gamma, \delta$ all converge a.s. to corresponding deterministic limits $\mathring{\Upsilon}, \mathring{\Lambda}, \mathring{\Gamma}, \mathring{\gamma}, \mathring{\delta}$:

$$\begin{aligned} \Upsilon_{ij} &\xrightarrow{\text{a.s.}} \mathring{\Upsilon}_{ij} \stackrel{\text{def}}{=} \mathbb{E}\, Z^{y^i}Z^{y^j} = (\sigma_A)^{-2}\operatorname{Cov}(\hat{Z}^{x^i}, \hat{Z}^{x^j}) \\ \Lambda_{ij} &\xrightarrow{\text{a.s.}} \mathring{\Lambda}_{ij} \stackrel{\text{def}}{=} \mathbb{E}\, Z^{v^i}Z^{v^j} = (\sigma_A)^{-2}\operatorname{Cov}(\hat{Z}^{u^i}, \hat{Z}^{u^j}) \\ \gamma_i &\xrightarrow{\text{a.s.}} \mathring{\gamma}_i \stackrel{\text{def}}{=} \mathbb{E}\, Z^{y^i}Z^h = (\sigma_A)^{-2}\operatorname{Cov}(\hat{Z}^{x^i}, \hat{Z}^g) \\ \Gamma_{ij} &\xrightarrow{\text{a.s.}} \mathring{\Gamma}_{ij} \stackrel{\text{def}}{=} \mathbb{E}\, Z^{u^i}Z^{y^j} \\ \delta_i &\xrightarrow{\text{a.s.}} \mathring{\delta}_i \stackrel{\text{def}}{=} \mathbb{E}\, Z^{u^i}Z^h \end{aligned} \tag{56}$$

It turns out that, as a consequence of Lemma L.11 below, we have: a.s. for all large enough $n$, $\operatorname{rank}\Upsilon = \operatorname{rank}\mathring{\Upsilon}$ and $\operatorname{rank}\Lambda = \operatorname{rank}\mathring{\Lambda}$. Therefore, as pseudoinverse is continuous on matrices of fixed rank, we get the following proposition

**Proposition L.5.** $\Upsilon^+ \xrightarrow{\text{a.s.}} \mathring{\Upsilon}^+$ *and* $\Lambda^+ \xrightarrow{\text{a.s.}} \mathring{\Lambda}^+$.

Using this proposition, we compute the limit of $\sigma^2$.

**Lemma L.6.** *The quantity $\sigma$, defined in Eq. (54), converges to a deterministic limit $\mathring{\sigma}$ almost surely:*

$$\sigma^2 \xrightarrow{\text{a.s.}} \mathring{\sigma}^2 \overset{\text{def}}{=} \sigma_A^2(\mathbb{E}(Z^h)^2 - \mathring{\gamma}^\top \mathring{\Upsilon}^+ \mathring{\gamma}). \tag{57}$$

*Proof.* Note that

$$\sigma^2 = \frac{\sigma_A^2}{n}(h^\top h - h^\top \mathbf{\Pi}_{\mathbf{Y}} h) = \frac{\sigma_A^2}{n}(h^\top h - h^\top \mathbf{Y}(\mathbf{Y}^\top \mathbf{Y})^+ \mathbf{Y}^\top h) = \frac{\sigma_A^2}{n}(h^\top h - \gamma^\top \Upsilon^+ \gamma).$$

Because $\phi$ is polynomially-bounded, so is $\phi(z)^2$. By induction hypothesis (Moments($m-1$)),

$$\frac{1}{n} h^\top h = \frac{1}{n}\sum_{\alpha=1}^n \phi(g_\alpha^1, \ldots, g_\alpha^{m-1})^2 \xrightarrow{\text{a.s.}} \mathbb{E}\,\phi(Z^{g^1}, \ldots, Z^{g^{m-1}})^2 = \mathbb{E}(Z^h)^2.$$

By Eq. (56), $\gamma \xrightarrow{\text{a.s.}} \mathring{\gamma}$ and $\Upsilon \xrightarrow{\text{a.s.}} \mathring{\Upsilon}$. By Proposition L.5, $\Upsilon^+ \xrightarrow{\text{a.s.}} \mathring{\Upsilon}^+$. Combining all of these limits together yields the desired claim. ☐

Using Eqs. (52) and (55), we can re-express $\omega$ as

$$\omega = \mathbf{X}\Upsilon^+\gamma + \mathbf{V}\Lambda^+\delta - \mathbf{V}\Lambda^+\Gamma\Upsilon^+\gamma.$$

Define $d \in \mathbb{R}^r$ and $e \in \mathbb{R}^s$ by

$$\begin{aligned} d &\overset{\text{def}}{=} \Upsilon^+\gamma, && \text{so that} \quad d \xrightarrow{\text{a.s.}} \mathring{d} \overset{\text{def}}{=} \mathring{\Upsilon}^+\mathring{\gamma}, \quad \text{and} \\ e &\overset{\text{def}}{=} \Lambda^+(\delta - \Gamma\Upsilon^+\gamma), && \text{so that} \quad e \xrightarrow{\text{a.s.}} \mathring{e} \overset{\text{def}}{=} \mathring{\Lambda}^+(\mathring{\delta} - \mathring{\Gamma}\mathring{\Upsilon}^+\mathring{\gamma}). \end{aligned} \tag{58}$$

Then the next lemma follows trivially.

**Lemma L.7.** *For $\omega$ defined in Eq. (54), we have*

$$\omega = Eh = \mathbf{X}d + \mathbf{V}e = \mathbf{X}(\mathring{d} + \hat{\varepsilon}) + \mathbf{V}(\mathring{e} + \check{\varepsilon}),$$

*for some random vectors $\hat{\varepsilon} \in \mathbb{R}^r, \check{\varepsilon} \in \mathbb{R}^s$ that go to 0 a.s. with $n$,*

### L.4.2 Helper Lemmas

Using the matrix notation of Definition L.1 and the expression for the projection operator, we can rewrite $\mathring{d}$ and $\mathring{e}$ as follows. The punchline of this section is Lemma L.10, which says

$$\sum_{i=1}^r \mathring{d}_i \dot{Z}^{x^i} + \sum_{j=1}^s \mathring{e}_j Z^{v^j} = \dot{Z}^g.$$

**Proposition L.8.** *Let $\Pi : \mathcal{L} \to \mathcal{L}$ be the orthogonal projection operator onto the linear span of $\{Z^{y^1}, \ldots, Z^{y^r}\}$. Then,*

$$\mathring{d} = \begin{bmatrix} & \mathring{\Upsilon}^+ & \end{bmatrix} \begin{bmatrix} Z^{y^1} \\ \vdots \\ Z^{y^r} \end{bmatrix} [Z^h], \quad \mathring{e} = \begin{bmatrix} & \mathring{\Lambda}^+ & \end{bmatrix} \begin{bmatrix} \hat{Z}^{u^1} \\ \vdots \\ \hat{Z}^{u^s} \end{bmatrix} [\Pi^\perp Z^h]. \tag{59}$$

*and*

$$\mathring{\sigma}^2 = \sigma_A^2\,\mathbb{E}\,Z^h \Pi^\perp Z^h = \sigma_A^2\,\mathbb{E}(\Pi^\perp Z^h)^2. \tag{60}$$

*Proof.* The identity for $\mathring{d}$ is obvious. We derive the identity for $\mathring{e}$, and the proof of Eq. (60) follows similarly. First, we expand the definitions of $\mathring{\delta}, \mathring{\Gamma}, \mathring{\Upsilon}, \mathring{\gamma}$ in matrix notation of Definition L.1.

$$\mathring{\delta} = \begin{bmatrix} Z^{u^1} \\ \vdots \\ Z^{u^s} \end{bmatrix} [Z^h], \quad \mathring{\Gamma} = \begin{bmatrix} Z^{u^1} \\ \vdots \\ Z^{u^s} \end{bmatrix} [Z^{y^1}, \ldots, Z^{y^r}], \quad \mathring{\Upsilon} = \begin{bmatrix} Z^{y^1} \\ \vdots \\ Z^{y^r} \end{bmatrix} [Z^{y^1}, \ldots, Z^{y^r}], \quad \mathring{\gamma} = \begin{bmatrix} Z^{y^1} \\ \vdots \\ Z^{y^r} \end{bmatrix} [Z^h].$$

Then with the matrix form of $\Pi$ in mind (Eq. (44)), we derive

$$\mathring{\delta} - \mathring{\Gamma}\mathring{\Upsilon}^+\mathring{\gamma} = \begin{bmatrix} Z^{u^1} \\ \vdots \\ Z^{u^s} \end{bmatrix} [(I - \Pi)Z^h] = \begin{bmatrix} Z^{u^1} \\ \vdots \\ Z^{u^s} \end{bmatrix} [\Pi^\perp Z^h] = \begin{bmatrix} \hat{Z}^{u^1} \\ \vdots \\ \hat{Z}^{u^s} \end{bmatrix} [\Pi^\perp Z^h]. \tag{61}$$

Here the last equality follows from the following

$$\begin{bmatrix} Z^{u^1} \\ \vdots \\ Z^{u^s} \end{bmatrix} [\Pi^\perp Z^h] = \begin{bmatrix} \Pi^\perp Z^{u^1} \\ \vdots \\ \Pi^\perp Z^{u^s} \end{bmatrix} [Z^h] = \begin{bmatrix} \Pi^\perp \hat{Z}^{u^1} \\ \vdots \\ \Pi^\perp \hat{Z}^{u^s} \end{bmatrix} [Z^h] = \begin{bmatrix} \hat{Z}^{u^1} \\ \vdots \\ \hat{Z}^{u^s} \end{bmatrix} [\Pi^\perp Z^h]$$

where the middle equality follows because each $Z^{u^j} - \hat{Z}^{u^j} = \dot{Z}^{u^j}$ is a linear combination of $\{Z^{y^i}\}_{i=1}^r$ by `ZDot`. Finally, plugging in Eq. (61) to Eq. (58), we have

$$\mathring{e} = \mathring{\Lambda}^+(\mathring{\delta} - \mathring{\Gamma}\mathring{\Upsilon}^+\mathring{\gamma}) = \begin{bmatrix} & \mathring{\Lambda}^+ & \end{bmatrix} \begin{bmatrix} \hat{Z}^{u^1} \\ \vdots \\ \hat{Z}^{u^s} \end{bmatrix} [\Pi^\perp Z^h].$$

□

Using Eq. (59), we can write the following

**Lemma L.9.** *Let $\Pi : \mathcal{L} \to \mathcal{L}$ be the orthogonal projection operator onto the linear span of $\{Z^{y^1}, \ldots, Z^{y^r}\}$. Then*

$$\sum_{i=1}^r \mathring{d}_i \dot{Z}^{x^i} = [Z^{v^1}, \ldots, Z^{v^s}] \begin{bmatrix} & \mathring{\Lambda}^+ & \end{bmatrix} \begin{bmatrix} \hat{Z}^{u^1} \\ \vdots \\ \hat{Z}^{u^s} \end{bmatrix} [\Pi Z^h]. \tag{62}$$

*Proof.* By Lemma L.4, for each $i \in [r]$,

$$\dot{Z}^{x^i} = [Z^{v^1}, \ldots, Z^{v^s}] \begin{bmatrix} & \mathring{\Lambda}^+ & \end{bmatrix} \begin{bmatrix} \hat{Z}^{u^1} \\ \vdots \\ \hat{Z}^{u^s} \end{bmatrix} [Z^{y^i}]$$

because $\{u^j\}_{j=1}^s$ contains all G-vars of the form $W^\top$ vector introduced before $x^i$. By Proposition L.8,

$$\mathring{d} = \begin{bmatrix} & \mathring{\Upsilon}^+ & \end{bmatrix} \begin{bmatrix} Z^{y^1} \\ \vdots \\ Z^{y^r} \end{bmatrix} [Z^h].$$

Then

$$
\begin{aligned}
\sum_{i=1}^{r} \mathring{d}_i \dot{Z}^{x^i} &= [\dot{Z}^{x^1}, \ldots, \dot{Z}^{x^r}] \begin{bmatrix} \\ \mathring{d} \\ \\ \end{bmatrix} \\
&= [Z^{v^1}, \ldots, Z^{v^s}] \begin{bmatrix} & & \\ & \mathring{\Lambda}^{+} & \\ & & \end{bmatrix} \begin{bmatrix} \hat{Z}^{u^1} \\ \vdots \\ \hat{Z}^{u^s} \end{bmatrix} [Z^{y^1}, \ldots, Z^{y^r}] \\
&\quad \times \begin{bmatrix} & & \\ & \mathring{\Upsilon}^{+} & \\ & & \end{bmatrix} \begin{bmatrix} Z^{y^1} \\ \vdots \\ Z^{y^r} \end{bmatrix} [Z^h] \\
&= (Z^{v^1}, \ldots, Z^{v^s}) \begin{bmatrix} & & \\ & \mathring{\Lambda}^{+} & \\ & & \end{bmatrix} \begin{bmatrix} \hat{Z}^{u^1} \\ \vdots \\ \hat{Z}^{u^s} \end{bmatrix} [\Pi Z^h]
\end{aligned}
$$

as desired. □

By combining Lemma L.9 with Proposition L.8, we get

**Lemma L.10.**

$$
\sum_{i=1}^{r} \mathring{d}_i \dot{Z}^{x^i} + \sum_{j=1}^{s} \mathring{e}_j Z^{v^j} = \dot{Z}^{g}.
$$

*Proof.* By realizing the $\Pi^{\perp}$ and $\Pi$ in Eq. (59) and Eq. (62) "cancel" to get identity, we have

$$
\begin{aligned}
\sum_{i=1}^{r} \mathring{d}_i \dot{Z}^{x^i} + \sum_{j=1}^{s} \mathring{e}_j Z^{v^j} &= [Z^{v^1}, \ldots, Z^{v^s}] \begin{bmatrix} & & \\ & \mathring{\Lambda}^{+} & \\ & & \end{bmatrix} \begin{bmatrix} \hat{Z}^{u^1} \\ \vdots \\ \hat{Z}^{u^s} \end{bmatrix} [Z^h] \\
&= [Z^{v^1}, \ldots, Z^{v^s}] \begin{bmatrix} & & \\ & C^{+} & \\ & & \end{bmatrix} \begin{bmatrix} \\ b \\ \\ \end{bmatrix}.
\end{aligned}
$$

where $C = \mathring{\Lambda}$ and $b$ are as in Eq. (4). By definition (`ZDot`), we have the desired result. □

### L.4.3 Index of Symbols

| | | |
|---|---|---|
| $g = g^m, h, \phi \cdots\cdots$ Eq. (48) | $x^i, y^i, u^j, v^j \cdots\cdots$ Eq. (49) | $\mathbf{X}, \mathbf{Y}, \mathbf{U}, \mathbf{V} \cdots\cdots$ Eq. (50) |
| $\mathbf{\Pi_V}, \mathbf{\Pi_Y} \cdots\cdots$ Eq. (51) | $E \cdots\cdots$ Eq. (52) | $\omega, \sigma \cdots\cdots$ Eq. (54) |
| $\Upsilon, \Lambda, \Gamma, \gamma, \delta \cdots\cdots$ Eq. (55) | $\mathring{\Upsilon}, \mathring{\Lambda}, \mathring{\Gamma}, \mathring{\gamma}, \mathring{\delta} \cdots\cdots$ Eq. (56) | $\mathring{\sigma} \cdots\cdots$ Eq. (57) |
| $d, e, \mathring{d}, \mathring{e} \cdots\cdots$ Eq. (58) | | |

In addition, $\mathcal{B}$ denotes the $\sigma$-algebra spanned by $g^1, \ldots, g^{m-1}$.

## L.5 Rank Stability and Zero Stability

In this section, we prove the following consequence of CoreSet$(m-1)$ and Moments$(m-1)$.

**Lemma L.11** (Rank Stability)**.** *For any collection of polynomially-bounded functions $\{\psi_j : \mathbb{R}^{m-1} \to \mathbb{R}\}_{j=1}^{l}$, let $K \in \mathbb{R}^{l \times l}$ be the random matrix (depending on $n$) defined by*

$$
K_{ij} = \frac{1}{n} \sum_{\alpha=1}^{n} \psi_i(g_\alpha^1, \ldots, g_\alpha^{m-1}) \psi_j(g_\alpha^1, \ldots, g_\alpha^{m-1}).
$$

*By Moments$(m-1)$,*

$$
K \xrightarrow{\text{a.s.}} \mathring{K}
$$

*for some matrix $\mathring{K} \in \mathbb{R}^{l \times l}$.*

1. *Then, almost surely, for large enough $n$,*

$$\ker K = \ker \mathring{K}, \quad \operatorname{im} K = \operatorname{im} \mathring{K}, \quad \textit{and} \quad \operatorname{rank} K = \operatorname{rank} \mathring{K}.$$

   *Here* $\ker$ *denotes null space and* $\operatorname{im}$ *denotes image space.*

2. *Suppose $I \subseteq [l]$ is any subset such that $\mathring{K}|_I$, the restriction of $\mathring{K}$ to rows and columns corresponding to $I$, satisfies*

$$|I| = \operatorname{rank} \mathring{K}|_I = \operatorname{rank} \mathring{K}.$$

   *There are unique coefficients $\{F_{ij}\}_{i\in[l],j\in I}$ that expresses each row of $\mathring{K}$ as linear combinations of rows corresponding to $I$:*

$$\forall i \in [l], \quad \mathring{K}_i = \sum_{j\in I} F_{ij} \mathring{K}_j.$$

   *Then, a.s. for all large $n$, for all $\alpha \in [n]$,*

$$\psi_i(g^1_\alpha, \dots, g^{m-1}_\alpha) = \sum_{j\in I} F_{ij} \psi_j(g^1_\alpha, \dots, g^{m-1}_\alpha).$$

Lemma L.11 will be primarily a corollary of the following Lemma L.12.

**Lemma L.12** (Zero Stability)**.** *If $\psi : \mathbb{R}^{m-1} \to \mathbb{R}^{\geq 0}$ is a nonnegative function such that*

$$\frac{1}{n}\sum_{\alpha=1}^{n} \psi(g^1_\alpha, \dots, g^{m-1}_\alpha) \xrightarrow{\text{a.s.}} 0$$

*then, almost surely, for large enough $n$,*

$$\psi(g^1_\alpha, \dots, g^{m-1}_\alpha) = 0$$

*for all $\alpha \in [n]$.*

We give the proof of Lemma L.11 now, assuming Lemma L.12.

*Proof.* Let $z \in \mathbb{R}^l$ be in the null space of $\mathring{K}$, i.e. $z^\top \mathring{K} z = 0$. Then we also have $z^\top K z \xrightarrow{\text{a.s.}} z^\top \mathring{K} z = 0$. But

$$z^\top K z = \frac{1}{n}\sum_{\alpha=1}^{n} \Psi(g^1_\alpha, \dots, g^{m-1}_\alpha), \quad \text{where} \quad \Psi(g^1_\alpha, \dots, g^{m-1}_\alpha) \stackrel{\text{def}}{=} \left(\sum_{i=1}^{l} z_i \psi_i(g^1_\alpha, \dots, g^{m-1}_\alpha)\right)^2$$

and $\Psi$ is a nonnegative function. By Lemma L.12, we have that: almost surely, for large enough $n$,

$$\Psi(g^1_\alpha, \dots, g^{m-1}_\alpha) = 0 \quad \text{for all } \alpha \in [n] \qquad \Longrightarrow z^\top K z = 0.$$

*Proof of Claim 1.* If we apply this argument to a basis $\{z^1, \dots, z^t\}$ of $\ker \mathring{K}$, then we get,

$$\text{a.s. for all large } n, \quad \ker \mathring{K} \subseteq \ker K,$$

so that

$$\text{a.s. for all large } n, \quad \operatorname{rank} \mathring{K} \geq \operatorname{rank} K.$$

Because the rank function is lower semicontinuous (i.e. the rank can drop suddenly, but cannot increase suddenly), and $K \xrightarrow{\text{a.s.}} \mathring{K}$, we also have

$$\text{a.s. for all large } n, \quad \operatorname{rank} \mathring{K} \leq \operatorname{rank} K.$$

Combined with the above, this gives the desired result on rank. The equality of null space then follows from the equality of rank, and the equality of image space follows immediately, as the image space is the orthogonal complement of the null space.

*Proof of Claim 2.* If we apply the above argument to each $z^i$ defined by inner product as

$$\forall x \in \mathbb{R}^l, \quad x^\top z^i = x_i - \sum_{j\in I} F_{ij} x_j,$$

(note that only for $i \notin I$ is $z^i$ nonzero), then we have, a.s. for large $n$, $z^{i\top} K z^i = 0$, or

$$\psi_i(g^1_\alpha, \dots, g^{m-1}_\alpha) = \sum_{j\in I} F_{ij} \psi_j(g^1_\alpha, \dots, g^{m-1}_\alpha), \quad \forall \alpha \in [n].$$

∎

In the rest of this section, we prove Lemma L.12. It helps to first show that the linear relations given in Basis carries over to the $n \to \infty$ limit.

**Proposition L.13.** *By the Basis property, each $g^i, i \in \mathcal{M} \subseteq [m-1]$, has a set of unique constants $\{a_j\}_{j \in \mathcal{M}}$ (independent of $n$) such that, almost surely, for large enough $n$,*

$$g^i = \sum_{j \in \mathcal{M}} a_j g^j.$$

*Then for each $i \in [m-1]$,*

$$Z^{g^i} \stackrel{\text{a.s.}}{=} \sum_{j \in \mathcal{M}} a_j Z^{g^j}.$$

*Proof.* Let $\psi(x^1, \ldots, x^{m-1}) \stackrel{\text{def}}{=} (x^i - \sum_{j \in \mathcal{M}} a_j x^j)^2$. Then by Basis$(m-1)$ and Moments$(m-1)$,

$$\frac{1}{n} \sum_{\alpha=1}^{n} \psi(g^1_\alpha, \ldots, g^{m-1}_\alpha) \xrightarrow{\text{a.s.}} \mathbb{E}\, \psi \left( Z^{g^1}, \ldots, Z^{g^{m-1}} \right) = 0.$$

This implies that

$$Z^{g^i} - \sum_{j \in \mathcal{M}} a_j Z^{g^j} \stackrel{\text{a.s.}}{=} 0 \iff Z^{g^i} \stackrel{\text{a.s.}}{=} \sum_{j \in \mathcal{M}} a_j Z^{g^j}$$

as desired. □

Now we show Lemma L.12.

*Proof of Lemma L.12.* By Moments$(m-1)$ and the premise of Lemma L.12,

$$\frac{1}{n} \sum_{\alpha=1}^{n} \psi(g^1_\alpha, \ldots, g^{m-1}_\alpha) \to \mathbb{E}\, \psi(Z^{g^1}, \ldots, Z^{g^{m-1}}) = 0.$$

Let $Z^{\mathcal{M}} \stackrel{\text{def}}{=} \{Z^{g^i}\}_{i \in \mathcal{M}}$. By Density$(m-1)$, the law of $Z^{\mathcal{M}}$ has density, i.e. a set is measure zero against its law iff it has measure zero against Lebesgue measure. Furthermore, Basis yields a linear function $F$ such that

$$\text{a.s. for large enough } n, \text{ for all } \alpha \in [n], \quad F(\{g^j_\alpha\}_{j \in \mathcal{M}}) = (g^1_\alpha, \ldots, g^{m-1}_\alpha).$$

By Proposition L.13, the same linear function satisfies

$$F(Z^{\mathcal{M}}) = (Z^{g^1}, \ldots, Z^{g^{m-1}}).$$

Therefore,

$$0 = \mathbb{E}\, \psi(Z^{g^1}, \ldots, Z^{g^{m-1}}) = \mathbb{E}\, \psi \circ F(Z^{\mathcal{M}}).$$

Because $\psi$, and thus $\psi \circ F$, is a nonnegative function, the nullity of the expectation implies that, other than a set $U$ of measure 0 under the distribution of $Z^{\mathcal{M}}$, $\psi \circ F$ is 0. This set $U$ also has Lebesgue measure zero as the law of $Z^{\mathcal{M}}$ has density, as discussed above.

If in NullAvoid$(m-1)$, we set $U_{n\alpha} = U$ for all $n$ and all $\alpha \in [n]$, then we get that: almost surely, for all large enough $n$, for all $\alpha \in [n]$,

$$\{g^i_\alpha\}_{i \in \mathcal{M}} \notin U \iff \psi \circ F(\{g^i_\alpha\}_{i \in \mathcal{M}}) = 0 \iff \psi(g^1_\alpha, \ldots, g^{m-1}_\alpha) = 0,$$

as desired. □

### L.6 Inductive Step: CoreSet$(m)$

In this section, we show

$$\text{Moments}(m-1) \text{ and } \text{CoreSet}(m-1) \implies \text{CoreSet}(m).$$

More explicitly, we need to think about whether to add $m$ to the core set $\mathcal{M}$ of $[m-1]$ in order to maintain the Basis, Density, and NullAvoid properties.

We proceed by casework on whether $\mathring{\sigma} = 0$.

### L.6.1 If $\mathring{\sigma} = 0$.

We will show that the core set properties are maintained if we don't add $m$ to the core set.

**$\mathcal{M}$ can be kept the same** Recall that $g = Ah$ and $x^i = Ay^i$ where $h$ was introduced by $h = \phi(g^1, \ldots, g^{m-1})$, for some polynomially bounded $\phi$. Suppose we likewise have $y^i = \hat{\phi}^i(g^1, \ldots, g^{m-1})$, for each $i \in [r]$, for polynomially bounded $\hat{\phi}^i : \mathbb{R}^{m-1} \to \mathbb{R}$ (where $\hat{\phi}^i$ only depends on the G-vars that came before $y^i$, but we implicitly pad coordinates of $\hat{\phi}^i$ to allow it to take all of $g^1, \ldots, g^{m-1}$ as inputs). By Basis, we know that, a.s. for large enough $n$, each of $g^1, \ldots, g^{m-1}$ is a (unique, constant-in-$n$) linear combination of $\{g^j\}_{j\in\mathcal{M}}$. Therefore, we can express

$$h = \psi(\{g^j\}_{j\in\mathcal{M}}), \quad \text{and} \quad \forall i \in [r], \quad y^i = \hat{\psi}^i(\{g^j\}_{j\in\mathcal{M}}) \tag{63}$$

for some polynomially bounded $\psi, \hat{\psi}^i : \mathbb{R}^{\mathcal{M}} \to \mathbb{R}$. Let $\Pi : \mathcal{L} \to \mathcal{L}$ be the orthogonal projection onto the subspace spanned by $Z^{y^1}, \ldots, Z^{y^r}$. By Eq. (60), we have $\mathring{\sigma}^2 = \sigma_A^2 \mathbb{E}(\Pi^\perp Z^h)^2$. Therefore, $\mathring{\sigma} = 0$ implies that

$$Z^h \overset{\text{a.s.}}{=} \Pi Z^h.$$

Since $Z^h, Z^{y^1}, \ldots, Z^{y^r}$ are resp. deterministic images of $Z^{\mathcal{M}}$ under the functions $\psi, \hat{\psi}^1, \ldots, \hat{\psi}^r$, this shows that $\psi$ is a linear combination of $\hat{\psi}^1, \ldots, \hat{\psi}^r$ almost surely under the law of $Z^{\mathcal{M}}$. But by Density$(m-1)$, the law of $Z^{\mathcal{M}}$ is absolutely continuous w.r.t. the Lesbegue measure (and vice versa), so this statement also holds under Lesbegue measure: For a set $U \subseteq \mathbb{R}^{\mathcal{M}}$ of Lesbegue measure zero and a set of coefficients $c_1, \ldots, c_r \in \mathbb{R}$, we have

$$\forall \vec{x} \notin U, \quad \psi(\vec{x}) = \sum_{i=1}^{r} c_i \hat{\psi}^i(\vec{x}).$$

By NullAvoid$(m-1)$ applied to $U_{n\alpha} = U$ for all $n$ and $\alpha \in [n]$, we also have that: a.s. for large enough $n$,

$$\psi(\{g^j\}_{j\in\mathcal{M}}) = \sum_{i=1}^{r} c_i \hat{\psi}^i(\{g^j\}_{j\in\mathcal{M}}), \quad \text{so that by Eq. (63)}$$

$$g = Ah = A\psi(\{g^j\}_{j\in\mathcal{M}}) = \sum_{i=1}^{r} c_i A\hat{\psi}^i(\{g^j\}_{j\in\mathcal{M}}), = \sum_{i=1}^{r} c_i A y^i = \sum_{i=1}^{r} c_i x^i.$$

This shows that, if we keep the core set as $\mathcal{M}$, then Basis is still satisfied. Since the core set is not changing, Density and NullAvoid just follows from the induction hypothesis. For usage later in the proof of Moments$(m)$, we record our observation here as a lemma.

**Lemma L.14.** *If $\mathring{\sigma} = 0$, then there are coefficients $c_1, \ldots, c_r \in \mathbb{R}$ independent of $n$ such that a.s. for large enough $n$,*

$$g = \sum_{i=1}^{r} c_i x^i.$$

### L.6.2 If $\mathring{\sigma} > 0$.

It's clear that $g$ cannot be in the linear span of $\{x^i\}_{i\in[r]}$ asymptotically, so we will add $g$ to the core set, and the Basis property follows immediately. In the below, we shall write $\mathcal{M}$ for the old core set, and $\mathcal{M}' \overset{\text{def}}{=} \mathcal{M} \cup \{g\}$ for the new one.

It remains to show Density and NullAvoid for $\mathcal{M}'$.

**Density$(m)$ holds** By definition (`ZMatMul`), we have

$$Z^g = \hat{Z}^g + \dot{Z}^g = \hat{Z}^g + \sum_{j=1}^{s} a_j Z^{v^j}$$

where $a_j$ are the partial derivative expectations in `ZDot` (whose specific values we will not care about) and Eq. (4). Note that for all $j \in [s]$, $Z^{v^j}$ only depends on $\hat{Z}^{g^1}, \ldots, \hat{Z}^{g^{m-1}}$. Let $\mathcal{Z}$ be the $\sigma$-algebra generated by $Z^{g^1}, \dots, Z^{g^{m-1}}$, which is the same as the $\sigma$-algebra generated by $\hat{Z}^{g^1}, \dots, \hat{Z}^{g^{m-1}}$ by Proposition I.7. Then conditioned on $\mathcal{Z}$, we can follow a quick calculation to see that

$$\hat{Z}^g \overset{\mathrm{d}}{=}_{\mathcal{Z}} \mathring{\sigma} z + \sum_{i=1}^{r} \mathring{d}_i \hat{Z}^{x^i}$$

where $z \sim \mathcal{N}(0,1)$ is independent from $\mathcal{Z}$, and $\mathring{d}$ is as in Lemma L.7. This allows us to write

$$Z^g \overset{\mathrm{d}}{=}_{\mathcal{Z}} \mathring{\sigma} z + F(\hat{Z}^{g^1}, \dots, \hat{Z}^{g^{m-1}})$$

for some deterministic function $F$. Since $\mathring{\sigma} > 0$, this shows that the distribution of $Z^g$ conditioned on $\mathcal{Z}$ is absolutely continuous w.r.t. the 1-dimensional Lebesgue measure, and vice versa. If we apply Lemma L.15 below with $X_1 = Z^g, X_2 = \{Z^{g^i}\}_{i \in \mathcal{M}}$, and $Y = (\hat{Z}^{g^1}, \dots, \hat{Z}^{g^{m-1}})$, then the lemma premise Eq. (65) follows from the above reasoning, and the lemma premise Eq. (64) follows from induction hypothesis $\text{Density}(m-1)$ (for $\mathcal{M}$). This then yields $\text{Density}(m)$ (for $\mathcal{M}'$), as desired.

**Lemma L.15.** *Consider a random vector $X = (X_1, X_2) \in \mathbb{R}^a \times \mathbb{R}^b$ and another random vector $Y \in \mathbb{R}^c$. Suppose $X_2$ is deterministic conditioned on $Y$. Let $\lambda$ denote the Lebesgue measure in any Euclidean space. If*

- *for every measurable set $U \subseteq \mathbb{R}^b$,*
$$\Pr(X_2 \in U) = 0 \iff \lambda(U) = 0, \quad \textit{and} \tag{64}$$
- *for every $y \in \mathbb{R}^c$, and every measurable set $U \subseteq \mathbb{R}^a$,*
$$\Pr(X_1 \in U \mid Y = y) = 0 \iff \lambda(U) = 0, \tag{65}$$

*then for every measurable set $V \subseteq \mathbb{R}^a \times \mathbb{R}^b$,*

$$\Pr(X \in V) = 0 \iff \lambda(V) = 0.$$

*Proof.* Fix $V \subseteq \mathbb{R}^a \times \mathbb{R}^b$. We have

$$\begin{aligned}\Pr(X \in V) &= \int \Pr(X \in V \mid Y = y) \,\mathrm{d}\Pr(Y = y) \\ &= \int \Pr(X_1 \in V_{x_2(y)} \mid Y = y) \,\mathrm{d}\Pr(Y = y),\end{aligned} \tag{66}$$

where $x_2(y)$ is the deterministic value of $X_2$ conditioned on $Y = y$, and $V_{x_2} = \{x_1 : (x_1, x_2) \in V\}$. Likewise,

$$\lambda(V) = \int \lambda(V_{x_2}) \,\mathrm{d}\lambda(x_2). \tag{67}$$

Then

$$\begin{aligned}
\Pr(X \in V) = 0 &\iff \Pr_Y(\Pr_{X_1}(X_1 \in V_{X_2} \mid Y) > 0) = 0 && \text{by Eq. (66)} \\
&\iff \Pr_Y(\lambda(V_{X_2}) > 0) = 0 && \text{by Eq. (65)} \\
&\iff \Pr_{X_2}(\lambda(V_{X_2}) > 0) = 0 && \text{because } X_2 \text{ is deterministic given } Y \\
&\iff \lambda(x_2 : \lambda(V_{x_2}) > 0) = 0 && \text{by Eq. (64)} \\
&\iff \lambda(V) = 0 && \text{by Eq. (67)}
\end{aligned}$$

as desired.

□

Now we tackle NullAvoid.

**NullAvoid**$(m)$ **holds** First, let's assume that, a.s. for large enough $n$, $\mathbf{\Pi}_{\mathbf{V}}^{\perp}$ has no zero diagonal entry; we shall show this fact below in Lemma L.16. Because the conditional variance of $g_{\alpha}^{m}$ given $g^1, \ldots, g^{m-1}$ is $\sigma^2(\mathbf{\Pi}_{\mathbf{V}}^{\perp})_{\alpha\alpha}$, and because $\mathring{\sigma} > 0$ by assumption in this section, this implies that, a.s. for all large enough $n$,

$$g_{\alpha}^{m}\text{, conditioned on } g^1, \ldots, g^{m-1}\text{, has density for all } \alpha \in [n]. \tag{68}$$

By "has density" here, we in particular mean that any Lesbegue measure zero set in $\mathbb{R}$ has zero probability under the conditional distribution of $g_{\alpha}^{m}$ given $g^1, \ldots, g^{m-1}$.

Now, assuming Lemma L.16, we prove NullAvoid holds for $\mathcal{M}'$.

Let $\{U_{n\alpha} \subseteq \mathbb{R}^{\mathcal{M}'}\}_{n\in\mathbb{N},\alpha\in[n]}$ be a triangular array of Lesbegue measure zero sets. For each $U_{n\alpha}$, define $B_{n\alpha} \stackrel{\text{def}}{=} \{\vec{x} \in \mathbb{R}^{\mathcal{M}} : \lambda(U_{n\alpha}|_{\vec{x}}) \neq 0\}$, where $U_{n\alpha}|_{\vec{x}} = \{y \in \mathbb{R} : (\vec{x}, y) \in U_{n\alpha} \subseteq \mathbb{R}^{\mathcal{M}} \times \mathbb{R}\}$ is the "slice" of $U_{n\alpha}$ at $\vec{x}$, and $\lambda$ is the 1-dimensional Lebesgue measure. Because each $U_{n\alpha}$ has measure zero in $\mathbb{R}^{\mathcal{M}'}$, necessarily each $B_{n\alpha}$ also has measure zero in $\mathbb{R}^{\mathcal{M}}$. Applying NullAvoid to the triangular array $\{B_{n\alpha} \subseteq \mathbb{R}^{\mathcal{M}}\}_{n\in\mathbb{N},\alpha\in[n]}$, we get that: a.s. for large enough $n$,

$$\forall \alpha \in [n], \quad \{g_{\alpha}^{i}\}_{i\in\mathcal{M}} \notin B_{n\alpha}.$$

Therefore, by Eq. (68), a.s. for large enough $n$,

$$\forall \alpha \in [n], \quad \{g_{\alpha}^{i}\}_{i\in\mathcal{M}'} \notin U_{n\alpha}.$$

This finishes the proof of NullAvoid for $\mathcal{M}'$, and also CoreSet$(m)$, save for Lemma L.16 below.

**Lemma L.16.** *Almost surely, for large enough $n$, $\mathbf{\Pi}_{\mathbf{V}}^{\perp}$ has no zero diagonal entry.*

*Proof.* WLOG, assume $\mathring{\Lambda}$ is full rank. Otherwise, by Lemma L.11(2), we can replace $v^1, \ldots, v^s$ by a linearly independent spanning set $v^{i_1}, \ldots, v^{i_k}$ such that 1) each $v^j$ is almost surely, for all large $n$, a linear combination of them and such that 2) their 2nd moment matrix is full rank in the limit. Then the projection matrix associated to $v^{i_1}, \ldots, v^{i_k}$ is, almost surely, for all large $n$, the same as $\mathbf{\Pi}_{\mathbf{V}}$.

By the Sherman-Morrison formula (Fact L.17),

$$(\mathbf{\Pi}_{\mathbf{V}})_{\alpha\alpha} = f\left(\frac{1}{n}\check{h}_{\alpha}{}^{\top}\Lambda_{-\alpha}^{-1}\check{h}_{\alpha}\right)$$

where $f(x) = x/(1+x)$, $\check{h}_{\alpha}$ is the column vector $(v_{\alpha}^{1}, \ldots, v_{\alpha}^{s})^{\top}$, and $\Lambda_{-\alpha} = \frac{1}{n}\sum_{\beta\neq\alpha}\check{h}_{\beta}\check{h}_{\beta}{}^{\top}$. Thus, unless $\Lambda_{-\alpha}$ is singular for some $\alpha$, all diagonal entries of $\mathbf{\Pi}_{\mathbf{V}}^{\perp} = I - \mathbf{\Pi}_{\mathbf{V}}$ are nonzero. So it suffices to show that,

$$\text{a.s. for large enough } n, \quad \Lambda_{-\alpha} \text{ is nonsingular for all } \alpha.$$

To do this, it pays to note that $\Lambda_{-\alpha} = \Lambda - \frac{1}{n}\check{h}_{\alpha}\check{h}_{\alpha}^{\top}$, so that

$$|\lambda_{\min}(\Lambda_{-\alpha}) - \lambda_{\min}(\Lambda)| \leq \|\frac{1}{n}\check{h}_{\alpha}\check{h}_{\alpha}^{\top}\|_{\text{op}} = \frac{1}{n}\check{h}_{\alpha}^{\top}\check{h}_{\alpha}.$$

By Lemma L.18 below (which bounds the max by a high moment),

$$\max_{\alpha\in n}\frac{1}{n}\check{h}_{\alpha}^{\top}\check{h}_{\alpha} \xrightarrow{\text{a.s.}} 0,$$

and consequently

$$\max_{\alpha\in[n]}|\lambda_{\min}(\Lambda_{-\alpha}) - \lambda_{\min}(\Lambda)| \xrightarrow{\text{a.s.}} 0.$$

Because $\Lambda \xrightarrow{\text{a.s.}} \mathring{\Lambda}$, we know that, a.s. for large enough $n$, $\lambda_{\min}(\Lambda)$ is bounded away from 0 by a constant (independent of $n$). Altogether, this implies that all $\Lambda_{-\alpha}$ are nonsingular, as desired. □

**Fact L.17** (Sherman-Morrison formula)**.** *For any nonsingular matrix $A \in \mathbb{R}^{l\times l}$ and vector $a \in \mathbb{R}^{l}$, we have*

$$a^{\top}(A + aa^{\top})^{-1}a = \frac{a^{\top}A^{-1}a}{1 + a^{\top}A^{-1}a}.$$

*Consequently, for any full rank matrix $H$, the $\alpha$th diagonal entry of its associated projection matrix $\Pi_H = H(H^\top H)^{-1}H^\top$ can be written as*

$$(\Pi_H)_{\alpha\alpha} = \frac{H_\alpha(H_{-\alpha}^\top H_{-\alpha})^{-1}H_\alpha^\top}{1 + H_\alpha(H_{-\alpha}^\top H_{-\alpha})^{-1}H_\alpha^\top}$$

*where $H_\alpha$ is the $\alpha$th row of $H$, and $H_{-\alpha}$ is $H$ with the $\alpha$th row removed.*

**Lemma L.18.** *Assume Moments$(m-1)$. Suppose $\psi : \mathbb{R}^{m-1} \to \mathbb{R}$ is polynomially bounded. Then as $n \to \infty$,*

$$\frac{1}{n^p} \max_{\alpha \in [n]} |\psi(g_\alpha^1, \ldots, g_\alpha^{m-1})| \xrightarrow{\text{a.s.}} 0$$

*for any $p > 0$.*

*Proof.* For any $q > 0$, we have the elementary bound

$$\max_{\alpha \in [n]} |\psi(g_\alpha^1, \ldots, g_\alpha^{m-1})| \leq \sqrt[q]{\sum_{\alpha \in [n]} |\psi(g_\alpha^1, \ldots, g_\alpha^{m-1})|^q}.$$

Thus, for any $q > 0$,

$$\frac{1}{n^p} \max_{\alpha \in [n]} |\psi(g_\alpha^1, \ldots, g_\alpha^{m-1})| \leq \frac{1}{n^{p-1/q}} \sqrt[q]{\frac{1}{n} \sum_{\alpha \in [n]} |\psi(g_\alpha^1, \ldots, g_\alpha^{m-1})|^q}.$$

Because, by Moments$(m-1)$, $\frac{1}{n}\sum_{\alpha \in [n]} |\psi(g_\alpha^1, \ldots, g_\alpha^{m-1})|^q \xrightarrow{\text{a.s.}} C$ for some finite constant $C$ as $n \to \infty$, the RHS above converges a.s. to 0 as soon as we take $q > 1/p$, and therefore so does the LHS.

$\square$

### L.7 Inductive Step: Moments$(m)$

In this section, we show

$$\text{Moments}(m-1) \text{ and } \text{CoreSet}(m-1) \implies \text{Moments}(m).$$

More specifically, we will show that for any polynomially-bounded $\psi : \mathbb{R}^m \to \mathbb{R}$,

$$\frac{1}{n} \sum_{\alpha=1}^{n} \psi(g_\alpha^1, \ldots, g_\alpha^m) \xrightarrow{\text{a.s.}} \mathbb{E}\, \psi(Z^{g^1}, \ldots, Z^{g^m}).$$

By Lemma L.14, if $\mathring{\sigma} = 0$, then almost surely, for large enough $n$, $g = g^m$ is just a (fixed) linear combination of $g^1, \ldots, g^{m-1}$, so Moments is trivially true. Therefore, in the below, we assume

$$\mathring{\sigma} > 0. \tag{$\star$}$$

This assumption will be crucial for our arguments involving smoothness induced by Gaussian averaging.

To clarify notation in the following, we will write $\mathbb{E}_X\,[expression]$ to denote the expectation over only the randomness in $X$, $\mathbb{E}\,[expression|\,\mathcal{B}]$ to denote the expectation taken over all randomness except those in $\mathcal{B}$, and $\mathbb{E}[expression]$ to denote expectation taken over *all randomness* in $expression$.

**Proof Plan** By triangle inequality, we decompose

$$\left| \frac{1}{n} \sum_{\alpha=1}^{n} \psi(g_\alpha^1, \ldots, g_\alpha^m) - \mathbb{E}\, \psi(Z^{g^1}, \ldots, Z^{g^m}) \right| \leq \mathsf{A} + \mathsf{B} + \mathsf{C} \tag{69}$$

where, with $\omega, \sigma$ (Eq. (54)), $\mathring{\sigma}$ (Eq. (57)), $\mathbf{\Pi_V}$ (Eq. (51)), $\mathring{d}, \mathring{e}$ (Eq. (58)) as in Appendix L.4, and with $z \sim \mathcal{N}(0,1)$, we have defined

$$\mathsf{A} \stackrel{\text{def}}{=} \left| \frac{1}{n} \sum_{\alpha=1}^{n} \psi(g_\alpha^1, \ldots, g_\alpha^m) - \mathbb{E}_z \psi\left( g_\alpha^1, \ldots, g_\alpha^{m-1}, \omega_\alpha + \sigma z \sqrt{(\mathbf{\Pi}_{\mathbf{V}}^{\perp})_{\alpha\alpha}} \right) \right|$$

$$\begin{aligned}\mathsf{B} \stackrel{\text{def}}{=} \Bigg| \frac{1}{n} \sum_{\alpha=1}^{n} & \mathbb{E}_z \psi\left( g_\alpha^1, \ldots, g_\alpha^{m-1}, \omega_\alpha + \sigma z \sqrt{(\mathbf{\Pi}_{\mathbf{V}}^{\perp})_{\alpha\alpha}} \right) \\ & - \mathbb{E}_z \psi\left( g_\alpha^1, \ldots, g_\alpha^{m-1}, \sum_{i=1}^{r} \mathring{d}_i x_\alpha^i + \sum_{j=1}^{s} \mathring{e}_j v_\alpha^j + \mathring{\sigma} z \right) \Bigg|\end{aligned}$$

$$\mathsf{C} \stackrel{\text{def}}{=} \left| \frac{1}{n} \sum_{\alpha=1}^{n} \mathbb{E}_z \psi\left( g_\alpha^1, \ldots, g_\alpha^{m-1}, \sum_{i=1}^{r} \mathring{d}_i x_\alpha^i + \sum_{j=1}^{s} \mathring{e}_j v_\alpha^j + \mathring{\sigma} z \right) - \mathbb{E} \psi(Z^{g^1}, \ldots, Z^{g^m}) \right|.$$

Note that $\mathsf{B}$ and $\mathsf{C}$ are random variables in $\mathcal{B}$, but $\mathsf{A}$ has additional randomness even after conditioning on $\mathcal{B}$. We will show that each of $\mathsf{A}, \mathsf{B}, \mathsf{C}$ goes to 0 almost surely, which would finish the proof of Theorem 2.10.

**High Level Logic** Roughly speaking,

$\mathsf{A} \xrightarrow{\text{a.s.}} 0$ because of a law of large numbers,

$\mathsf{B} \xrightarrow{\text{a.s.}} 0$ because of the smoothness in $\mathbb{E}_z \psi$ induced by Gaussian averaging, and

$\mathsf{C} \xrightarrow{\text{a.s.}} 0$ by induction hypothesis.

We start by proving $\mathsf{C} \xrightarrow{\text{a.s.}} 0$, since it's the easiest.

### L.7.1 $\mathsf{C}$ Converges Almost Surely to 0

In this section we show that $\mathsf{C} \xrightarrow{\text{a.s.}} 0$ by a straightforward reduction to the induction hypothesis.

A simple inductive argument with Box 1 shows that $Z^{x^1}, \ldots, Z^{x^r}, Z^{v^1}, \ldots, Z^{v^s}$ are all deterministic, polynomially-bounded functions of $Z^{g^1}, \ldots, Z^{g^{m-1}}$. Thus we may define the function $\Psi : \mathbb{R}^{m-1} \to \mathbb{R}$ by

$$\Psi(Z^{g^1}, \ldots, Z^{g^{m-1}}) \stackrel{\text{def}}{=} \mathbb{E}_{z \sim \mathcal{N}(0,1)} \psi\left( Z^{g^1}, \ldots, Z^{g^{m-1}}, \sum_{i=1}^{r} \mathring{d}_i Z^{x^i} + \sum_{j=1}^{s} \mathring{e}_j Z^{v^j} + \mathring{\sigma} z \right).$$

The function $\Psi$ is polynomially bounded since $\psi$ is, and $Z^{x^1}, \ldots, Z^{x^r}, Z^{v^1}, \ldots, Z^{v^s}$ are polynomially-bounded functions of $Z^{g^1}, \ldots, Z^{g^{m-1}}$. Observe that the function that expresses $Z^{x^i}$ in terms of $Z^{g^1}, \ldots, Z^{g^{m-1}}$ is the same function that expresses $x_\alpha^i$ in terms of $g_\alpha^1, \ldots, g_\alpha^{m-1}$ for all $\alpha \in [n]$; likewise for $Z^{v^j}$ and $v_\alpha^j$. Applying the induction hypothesis Moments$(m-1)$ to $\Psi$, we obtain

$$\begin{aligned}& \frac{1}{n} \sum_{\alpha=1}^{n} \mathbb{E}_z \psi\left( g_\alpha^1, \ldots, g_\alpha^{m-1}, \sum_{i=1}^{r} \mathring{d}_i x_\alpha^i + \sum_{j=1}^{s} \mathring{e}_j v_\alpha^j + \mathring{\sigma} z \right) \\ & = \frac{1}{n} \sum_{\alpha=1}^{n} \Psi\left( g_\alpha^1, \ldots, g_\alpha^{m-1} \right) \\ & \xrightarrow{\text{a.s.}} \mathbb{E} \Psi(Z^{g^1}, \ldots, Z^{g^{m-1}}) \\ & \qquad \qquad \qquad \qquad \text{by Moments}(m-1) \\ & = \mathbb{E} \psi\left( Z^{g^1}, \ldots, Z^{g^{m-1}}, \sum_{i=1}^{r} \mathring{d}_i Z^{x^i} + \sum_{j=1}^{s} \mathring{e}_j Z^{v^j} + \mathring{\sigma} z \right).\end{aligned}$$

By Lemma L.19 below, this is precisely

$$\mathbb{E}\,\psi\left(Z^{g^1},\ldots,Z^{g^{m-1}},Z^{g^m}\right)$$

as desired.

**Lemma L.19.** *Let $\mathcal{Z}$ be the $\sigma$-algebra generated by $Z^{g^1},\ldots,Z^{g^{m-1}}$. Then for $z\sim\mathcal{N}(0,1)$ sampled independently of $\mathcal{Z}$,*

$$Z^g \overset{\mathrm{d}}{=}_{\mathcal{Z}} \mathring{\sigma} z+\sum_{i=1}^{r}\mathring{d}_i Z^{x^i}+\sum_{j=1}^{s}\mathring{e}_j Z^{v^j}$$

*where $\mathring{d}$ and $\mathring{e}$ are as in Eq. (58) and $\mathring{\sigma}$ is as in Eq. (57).*

*Proof.* We can split each $Z^{x^i}$ into $\hat{Z}^{x^i}+\dot{Z}^{x^i}$, to obtain

$$\begin{aligned}
&\mathring{\sigma} z+\sum_{i=1}^{r}\mathring{d}_i Z^{x^i}+\sum_{j=1}^{s}\mathring{e}_j Z^{v^j}\\
&=\left(\mathring{\sigma} z+\sum_{i=1}^{r}\mathring{d}_i \hat{Z}^{x^i}\right)+\left(\sum_{i=1}^{r}\mathring{d}_i \dot{Z}^{x^i}+\sum_{j=1}^{s}\mathring{e}_j Z^{v^j}\right)\\
&=\left(\mathring{\sigma} z+\sum_{i=1}^{r}\mathring{d}_i \hat{Z}^{x^i}\right)+\dot{Z}^g && \text{by Lemma L.10}\\
&\overset{\mathrm{d}}{=}_{\mathcal{Z}} \hat{Z}^g+\dot{Z}^g && \text{see below for justification}\\
&= Z^g,
\end{aligned}$$

as desired. It remains to justify the second-to-last equality, which is easily done via the usual formula for Gaussian conditioning (Proposition K.3): Let $\hat{Z}$ be the column vector $(\hat{Z}^{x^1},\ldots,\hat{Z}^{x^r})^\top$. Then we know that $\hat{Z}$ and $\hat{Z}^g$ are jointly Gaussian with zero mean[47]. The covariance between $\hat{Z}$ and $\hat{Z}^g$ is $\sigma_A^2\mathring{\gamma}$ and $\hat{Z}$ has covariance matrix $\sigma_A^2\mathring{\Upsilon}$ by Eq. (56). Then by Proposition K.3, we know that, conditioned on $\mathcal{Z}$, $\hat{Z}^g$ is distributed as a Gaussian with mean

$$\sigma_A^2\mathring{\gamma}^\top(\sigma_A^2\mathring{\Upsilon})^+\hat{Z}=\mathring{d}^\top\hat{Z}=\sum_{i=1}^{r}\mathring{d}_i\hat{Z}^{x^i},$$

by Eq. (58) and variance

$$\mathbb{E}(Z^g)^2-\sigma_A^2\mathring{\gamma}^\top(\sigma_A^2\mathring{\Upsilon})^+\sigma_A^2\mathring{\gamma}=\mathring{\sigma}$$

by Lemma L.6. This is exactly what we needed. ☐

#### L.7.2 $\mathsf{A}$ Converges Almost Surely to 0

In this section we show $\mathsf{A}\xrightarrow{\text{a.s.}}0$.

For each $\alpha\in[n]$, define the function $\psi_\alpha:\mathbb{R}\to\mathbb{R}$ by

$$\psi_\alpha(x)\overset{\text{def}}{=}\psi(g^1_\alpha,\ldots,g^{m-1}_\alpha,\omega_\alpha+\sigma x),$$

with $\omega$ and $\sigma$ defined in Eq. (54). This is a random function depending on the random vectors $g^1_\alpha,\ldots,g^{m-1}_\alpha$, and it changes with $n$ as well. We also consider the "centered version" $\tilde{\psi}_\alpha:\mathbb{R}\to\mathbb{R}$ of $\psi_\alpha$,

$$\tilde{\psi}_\alpha(x)\overset{\text{def}}{=}\psi_\alpha(x)-\mathop{\mathbb{E}}_{y}\psi_\alpha(y)$$

with expectation taken over $y\sim\mathcal{N}(0,(\mathbf{\Pi}^\perp_{\mathbf{V}})_{\alpha\alpha})$ (but not $g^1_\alpha,\ldots,g^{m-1}_\alpha$). Note by Eq. (53),

$$\mathsf{A}\overset{\mathrm{d}}{=}_{\mathcal{B}}\left|\frac{1}{n}\sum_{\alpha=1}^{n}\tilde{\psi}_\alpha(z_\alpha)\right|$$

where $z\sim\mathcal{N}(0,\mathbf{\Pi}^\perp_{\mathbf{V}})$.

[47] recall we have assumed WLOG that $\mathbb{E}\,Z^g=0$ for all $g\in\mathcal{V}$; see discussion at the beginning of Appendix L.

**Proof idea** To prove our claim, we will show that, for almost all (i.e. probability 1 in the probability space U defined in the beginning of Appendix L) sequences of $(g^1, \ldots, g^{m-1}) = (g^1(n), \ldots, g^{m-1}(n))$ in $n$ — which we shall call *amenable sequences of* $g^1, \ldots, g^{m-1}$ — we have a moment bound

$$\mathbb{E}[\mathsf{A}^{2\lambda}|\mathcal{B}] = \mathop{\mathbb{E}}_{z\sim\mathcal{N}(0,\mathbf{\Pi}_{\mathbf{V}}^{\perp})} \left(\frac{1}{n}\sum_{\alpha=1}^{n} \tilde{\psi}_\alpha(z_\alpha)\right)^{2\lambda} < Cn^{-1.25} \tag{70}$$

for some large $\lambda$ and some constant $C > 0$ depending only on $\lambda$ and the particular sequence of $\{(g^1(n), \ldots, g^{m-1}(n))\}_n$. Then we apply Lemma K.2 to show that, conditioned on any amenable sequence, A converges to 0 almost surely over all randomness remaining after conditioning. Since almost all sequences are amenable, this shows that the convergence is also almost sure without the conditioning.

**The moment bound** For $\lambda \geq 6$ and any $q > 1$, we first apply Theorem K.23 to get the bound

$$\mathop{\mathbb{E}}_{z} \left(\frac{1}{n}\sum_{\alpha=1}^{n} \tilde{\psi}_\alpha(z_\alpha)\right)^{2\lambda} \leq cn^{-1.5+1/q} \sqrt[q]{\frac{1}{n}\sum_{\alpha=1}^{n} \mathbb{E}\,\tilde{\psi}_\alpha(z_\alpha)^{2\lambda q}}$$

where on both sides $z \sim \mathcal{N}(0, \mathbf{\Pi}_{\mathbf{V}}^{\perp})$, and $c$ is a constant depending only on $\lambda$ and $m$, but not on $n$, the functions $\psi_\alpha$, or $g^1, \ldots, g^{m-1}$. To obtain Eq. (70), we will show that

$$\frac{1}{n}\sum_{\alpha=1}^{n} \mathbb{E}\,\tilde{\psi}_\alpha(z_\alpha)^{2\lambda q} \tag{71}$$

is uniformly bounded (in $n$), almost surely over the randomness of the sequences $\{g^1(n), \ldots, g^{m-1}(n)\}_n$. We take all such sequences to be the *amenable sequences*. For $q > 4$, we then get the desired moment bound Eq. (70).

It remains to show the almost sure uniform boundedness.

**Almost sure uniform boundedness** Intuitively, Eq. (71) should converge almost surely to a deterministic value by applying some version of the induction hypothesis, so it should be almost surely uniformly bounded in $n$. The obstacle is that the variance $(\mathbf{\Pi}_{\mathbf{V}}^{\perp})_{\alpha\alpha}$ of $z_\alpha$ depends not only on $g^1_\alpha, \ldots, g^{m-1}_\alpha$, but also on $g^1_\beta, \ldots, g^{m-1}_\beta$ for other indices $\beta \neq \alpha$ as well. So *a priori* it is not clear how to apply the Moments$(m-1)$ in a straightforward way. We thus first process Eq. (71) a bit. Let

$$\mu_\alpha \stackrel{\text{def}}{=} \mathop{\mathbb{E}}_{x\sim\mathcal{N}(0,(\mathbf{\Pi}_{\mathbf{V}}^{\perp})_{\alpha\alpha})} \psi_\alpha(x).$$

Then, abbreviating $\mathbb{E}_z$ for expectation taken over $z \sim \mathcal{N}(0, \mathbf{\Pi}_{\mathbf{V}}^{\perp})$, we have the following inequalities of random variables in $\mathcal{B}$:

$$\begin{aligned}
\frac{1}{n}\sum_{\alpha=1}^{n} \mathop{\mathbb{E}}_{z}\tilde{\psi}_\alpha(z_\alpha)^{2\lambda q} &= \frac{1}{n}\sum_{\alpha=1}^{n} \mathop{\mathbb{E}}_{z}(\psi_\alpha(z_\alpha) - \mu_\alpha)^{2\lambda q} \\
&\leq \frac{1}{n}2^{2\lambda q-1}\sum_{\alpha=1}^{n} \mathop{\mathbb{E}}_{z}\left[\psi_\alpha(z_\alpha)^{2\lambda q} + \mu_\alpha^{2\lambda q}\right] && \text{by Lemma K.1} \\
&\leq \frac{1}{n}2^{2\lambda q}\sum_{\alpha=1}^{n} \mathop{\mathbb{E}}_{z}\psi_\alpha(z_\alpha)^{2\lambda q} && \text{see below} \\
&= \frac{1}{n}2^{2\lambda q}\sum_{\alpha=1}^{n} \mathop{\mathbb{E}}_{z}\psi(g^1_\alpha, \ldots, g^{m-1}_\alpha, \omega_\alpha + \sigma z_\alpha)^{2\lambda q},
\end{aligned}$$

where in the second inequality we applied power mean inequality $\mu_\alpha \leq \sqrt[2\lambda q]{\mathbb{E}_z\,\psi_\alpha(z_\alpha)^{2\lambda q}}$. Suppose, WLOG, that $\psi$ is polynomially bounded by an inequality $|\psi(x)| \leq C\|x\|_p^p + c$ for some $p, C, c > 0$. In the below, we will silently introduce constants $C_1, C_2, \ldots$ via Lemma K.1 and merge with old constants, such that they will only depend on $\lambda, p, q$. Continuing the chain of inequalities above

$$\begin{aligned}\frac{1}{n}\sum_{\alpha=1}^{n}\mathop{\mathbb{E}}_{z}\tilde{\psi}_\alpha(z_\alpha)^{2\lambda q} &\leq c + \frac{1}{n}C2^{2\lambda q}\sum_{\alpha=1}^{n}\mathop{\mathbb{E}}_{z}\left(|g^1_\alpha|^p + \ldots + |g^{m-1}_\alpha|^p + |\omega_\alpha + \sigma z_\alpha|^p\right)^{2\lambda q}\\ &\leq c + \frac{1}{n}C_1\sum_{\alpha=1}^{n}\mathop{\mathbb{E}}_{z}\left(|g^1_\alpha|^p + \ldots + |g^{m-1}_\alpha|^p + |\omega_\alpha|^p + |\sigma z_\alpha|^p\right)^{2\lambda q}\\ &\leq c + \frac{1}{n}C_2\sum_{\alpha=1}^{n}\mathop{\mathbb{E}}_{z}|g^1_\alpha|^{2\lambda qp} + \ldots + |g^{m-1}_\alpha|^{2\lambda qp} + |\omega_\alpha|^{2\lambda qp} + |\sigma z_\alpha|^{2\lambda qp}.\end{aligned} \tag{72}$$

We now proceed to show that the summands of Eq. (72) are almost surely uniformly bounded, which finishes our proof of $\mathsf{A} \xrightarrow{\text{a.s.}} 0$.

- By Moments$(m-1)$,

$$\frac{1}{n}\sum_{\alpha=1}^{n}\mathop{\mathbb{E}}_{z}|g^1_\alpha|^{2\lambda qp} + \ldots + |g^{m-1}_\alpha|^{2\lambda qp}$$

almost surely converges to a deterministic value, so it is almost surely uniformly bounded in $n$.

- In addition, $\sigma \xrightarrow{\text{a.s.}} \mathring{\sigma}$, so that, almost surely, for large enough $n$, we have $\sigma \leq \mathring{\sigma} + 1$. (The order of the qualifiers is important here; in general this statement cannot be made uniformly in $n$). Therefore, *almost surely, for large enough* $n$,

$$\frac{1}{n}\sum_{\alpha=1}^{n}\mathop{\mathbb{E}}_{z}|\sigma z_\alpha|^{2\lambda qp} \leq \frac{1}{n}\sum_{\alpha=1}^{n}|\mathring{\sigma}+1|^{2\lambda qp}\mathop{\mathbb{E}}_{z}|z_\alpha|^{2\lambda qp}.$$

This is almost surely uniformly bounded in $n$ because $\mathrm{Var}(z_\alpha) = (\mathbf{\Pi}^{\perp}_{\mathbf{V}})_{\alpha\alpha} \in [0,1]$ for all $\alpha$ by Lemma K.18, so that $\mathbb{E}_z |z_\alpha|^{2\lambda qp}$ (which is purely a monotonic function of $\mathrm{Var}(z_\alpha)$) is bounded as well.

- It remains to bound $\frac{1}{n}\sum_{\alpha=1}^{n}|\omega_\alpha|^{2\lambda qp}$. We extract our reasoning here into the Lemma L.20 below, as we will need to reuse this for later. This finishes the proof of $\mathsf{A} \xrightarrow{\text{a.s.}} 0$.

**Lemma L.20.** *For any polynomially bounded function* $\varphi : \mathbb{R} \to \mathbb{R}$,

$$\frac{1}{n}\sum_{\alpha=1}^{n}|\varphi(\omega_\alpha)|$$

*is almost surely uniformly bounded in* $n$.

*Proof.* It suffices to prove this for $\varphi(x) = |x|^d$ for any $d > 0$.

Expanding $\omega$ according to Lemma L.7, we get

$$\frac{1}{n}\sum_{\alpha=1}^{n}|\omega_\alpha|^d = \frac{1}{n}\sum_{\alpha=1}^{n}\left|\sum_{i=1}^{r}x^i_\alpha(\mathring{d}_i + \hat{\varepsilon}_i) + \sum_{j=1}^{s}v^j_\alpha(\mathring{e}_j + \check{\varepsilon}_j)\right|^d$$

for (fixed dimensional) $\hat{\varepsilon} \in \mathbb{R}^r, \check{\varepsilon} \in \mathbb{R}^s$ that go to 0 almost surely with $n$. Applying Lemma K.1, we get

$$\frac{1}{n}\sum_{\alpha=1}^{n}|\omega_\alpha|^d \leq \frac{1}{n}C_3\sum_{\alpha=1}^{n}\left|\sum_{i=1}^{r}x^i_\alpha\mathring{d}_i\right|^d + \left|\sum_{j=1}^{s}v^j_\alpha\mathring{e}_i\right|^d + \left|\sum_{i=1}^{r}x^i_\alpha\hat{\varepsilon}_i\right|^d + \left|\sum_{j=1}^{s}v^j_\alpha\check{\varepsilon}_j\right|^d.$$

We bound each summand separately.

- By induction hypothesis,

$$\frac{1}{n}\sum_{\alpha=1}^{n}\left|\sum_{i=1}^{r} x_\alpha^i \mathring{d}_i\right|^d$$

converges a.s. to a deterministic value, so it is a.s uniformly bounded in $n$.

- By the a.s. decaying property of $\hat{\varepsilon}$, we have almost surely, for large enough $n$, $|\sum_{i=1}^r x_\alpha^i \hat{\varepsilon}_i| \le \sum_{i=1}^r |x_\alpha^i|$ (again, the order of qualifier is very important here). By induction hypothesis,

$$\frac{1}{n}\sum_{\alpha=1}^{n}\left(\sum_{i=1}^{r} |x_\alpha^i|\right)^d$$

converges a.s. to a deterministic value, yielding the a.s. uniform-boundedness of it and of

$$\frac{1}{n}\sum_{\alpha=1}^{n}\left|\sum_{i=1}^{r} x_\alpha^i \hat{\varepsilon}_i\right|^d .$$

- Likewise, because for each $j$, $v^j$ is a polynomially-bounded function of $g^1, \ldots, g^{m-1}$ [48], the summands of

$$\frac{1}{n}\sum_{\alpha=1}^{n}\left(\sum_{j=1}^{s} |v_\alpha^j|\right)^d \quad \text{and} \quad \frac{1}{n}\sum_{\alpha=1}^{n}\left|\sum_{j=1}^{s} v_\alpha^j \mathring{e}_j\right|^d$$

are polynomially-bounded functions of $g^1, \ldots, g^{m-1}$ too. So by induction hypothesis, these sums converge a.s., implying the a.s. uniform-boundedness of them and of

$$\frac{1}{n}\sum_{\alpha=1}^{n}\left|\sum_{j=1}^{s} v_\alpha^j \check{\varepsilon}_j\right|^d .$$

□

#### L.7.3 $\mathsf{B}$ Converges Almost Surely to 0

In this section we show $\mathsf{B} \xrightarrow{\text{a.s.}} 0$.

**Some Notations** For brevity, we will set $d_\alpha \stackrel{\text{def}}{=} (\mathbf{\Pi}_{\mathbf{V}}^{\perp})_{\alpha\alpha}$. In addition, for each $\alpha \in [n]$, $w \in \mathbb{R}$, $\tau \ge 0$, define the function $\Psi_\alpha(-;-) : \mathbb{R} \times \mathbb{R}^{\ge 0} \to \mathbb{R}$ by

$$\Psi_\alpha(w; \tau^2) \stackrel{\text{def}}{=} \mathop{\mathbb{E}}_{z \sim \mathcal{N}(0,1)} \psi\left(g_\alpha^1, \ldots, g_\alpha^{m-1}, w + \tau z\right).$$

(Here and in all that follows, $\tau^2$ is the square of $\tau$, and the 2 is not an index). This is a random function, with randomness induced by $g^1, \ldots, g^{m-1}$.

**Our proof idea** is to write

$$\begin{aligned}
\mathsf{B} &= \left|\frac{1}{n}\sum_{\alpha=1}^{n} \Psi_\alpha\left(\omega_\alpha; \sigma^2 d_\alpha\right) - \Psi_\alpha\left(\sum_{i=1}^{r} \mathring{d}_i x_\alpha^i + \sum_{j=1}^{s} \mathring{e}_j v_\alpha^j; \mathring{\sigma}^2\right)\right| \\
&\le \frac{1}{n}\sum_{\alpha \in \bar{J}} \left|\Psi_\alpha\left(\omega_\alpha; \sigma^2 d_\alpha\right)\right| + \left|\Psi_\alpha\left(\sum_{i=1}^{r} \mathring{d}_i x_\alpha^i + \sum_{j=1}^{s} \mathring{e}_j v_\alpha^j; \mathring{\sigma}^2\right)\right| && (73) \\
&\quad + \frac{1}{n}\sum_{\alpha \in J} \left|\Psi_\alpha\left(\omega_\alpha; \sigma^2 d_\alpha\right) - \Psi_\alpha\left(\sum_{i=1}^{r} \mathring{d}_i x_\alpha^i + \sum_{j=1}^{s} \mathring{e}_j v_\alpha^j; \mathring{\sigma}^2\right)\right| && (74)
\end{aligned}$$

[48] This is the most crucial place where we need the assumption that all nonlinearities in the program are polynomially bounded. Otherwise, for faster growing functions, the compositions of such nonlinearities might not be integrable against the Gaussian measure

where $J \sqcup \bar{J} = [n]$ is a partition of $[n]$ with $J \overset{\text{def}}{=} \{\alpha : d_\alpha \geq 1/2\}$ and $\bar{J}$ is its complement. Note that $|\bar{J}| \leq 2 \operatorname{rank} \mathbf{V} \leq 2s$ is uniformly bounded in $n$. We then show each summand of Eq. (73) goes to 0 a.s. individually. Finally we use the smoothness of $\Psi_\alpha$ (Eq. (76)) induced by the Gaussian averaging in $z$ to show each summand of Eq. (74) is almost surely $o(1)$, finishing the proof.

**Eq. (73) converges to 0 a.s.** We first look at the term

$$\frac{1}{n}\sum_{\alpha \in \bar{J}} \left| \Psi_\alpha \left( \sum_{i=1}^r \mathring{d}_i x_\alpha^i + \sum_{j=1}^s \mathring{e}_j v_\alpha^j ; \mathring{\sigma}^2 \right) \right| \leq \frac{|\bar{J}|}{n} \max_{\alpha \in [n]} \left| \Psi_\alpha \left( \sum_{i=1}^r \mathring{d}_i x_\alpha^i + \sum_{j=1}^s \mathring{e}_j v_\alpha^j ; \mathring{\sigma}^2 \right) \right|$$
$$\leq \frac{2s}{n^{1-1/q}} \sqrt[q]{\frac{1}{n}\sum_{\alpha \in [n]} \left| \Psi_\alpha \left( \sum_{i=1}^r \mathring{d}_i x_\alpha^i + \sum_{j=1}^s \mathring{e}_j v_\alpha^j ; \mathring{\sigma}^2 \right) \right|^q} \tag{75}$$

for any $q > 0$. Here we used $|\bar{J}| \leq 2 \operatorname{rank} \mathbf{V} \leq 2s$ as noted above. Now $\left| \Psi_\alpha \left( \sum_{i=1}^r \mathring{d}_i x_\alpha^i + \sum_{j=1}^s \mathring{e}_j v_\alpha^j ; \mathring{\sigma}^2 \right) \right|^q$ is a fixed (independent of $\alpha$ and $n$) polynomially-bounded function of $g_\alpha^1, \ldots, g_\alpha^{m-1}$, so by induction hypothesis,

$$\frac{1}{n}\sum_{\alpha \in [n]} \left| \Psi_\alpha \left( \sum_{i=1}^r \mathring{d}_i x_\alpha^i + \sum_{j=1}^s \mathring{e}_j v_\alpha^j ; \mathring{\sigma}^2 \right) \right|^q$$

is a.s. uniformly bounded in $n$, so that using a large $q \geq 2$, we see Eq. (75) converges a.s. to 0.

Next, we apply a similar reasoning to the other term and obtain

$$\frac{1}{n}\sum_{\alpha \in \bar{J}} \left| \Psi_\alpha \left( \omega_\alpha ; \sigma^2 d_\alpha \right) \right| \leq \frac{2s}{n^{1-1/q}} \sqrt[q]{\frac{1}{n}\sum_{\alpha \in [n]} \left| \Psi_\alpha \left( \omega_\alpha ; \sigma^2 d_\alpha \right) \right|^q}$$

We in fact already know that

$$\frac{1}{n}\sum_{\alpha \in [n]} \left| \Psi_\alpha \left( \omega_\alpha ; \sigma^2 d_\alpha \right) \right|^q$$

is a.s. uniformly bounded in $n$ from Eq. (72) in Appendix L.7.2, so that

$$\frac{1}{n}\sum_{\alpha \in \bar{J}} \left| \Psi_\alpha \left( \omega_\alpha ; \sigma^2 d_\alpha \right) \right| \xrightarrow{\text{a.s.}} 0$$

from which follows the same for Eq. (73).

**Eq. (74) converges to 0 a.s.** As mentioned above, to prove this we will use the following smoothness bound of $\Psi_\alpha$, whose proof will be delayed to the end of the section. Suppose, WLOG, that the polynomially boundedness of $\psi$ presents itself in an inequality $|\psi(x)| \leq C\|x\|_p^p + C$, for some $p, C > 0$, where $p$ is an integer. This $p$ will appear explicitly in this smoothness bound below.

**Lemma L.21** (Smoothness of $\Psi_\alpha$)**.** *Let $w, \Delta w \in \mathbb{R}, \tau^2, \Delta\tau^2 \in \mathbb{R}^{\geq 0}$. Then*

$$\left| \Psi_\alpha(w + \Delta w; \tau^2 + \Delta\tau^2) - \Psi_\alpha(w; \tau^2) \right|$$
$$\leq R(|\Delta w| + \Delta\tau^2)(1 + \tau^{-2}) \left( S_\alpha + |w|^p + |\Delta w|^p + \tau^p + (\Delta\tau^2)^{p/2} \right) \tag{76}$$

*for some constant $R > 0$, and where*

$$S_\alpha \overset{\text{def}}{=} 1 + |g_\alpha^1|^p + \cdots + |g_\alpha^{m-1}|^p.$$

To bound Eq. (74), first we expand

$$\omega_\alpha = \sum_{i=1}^r x_\alpha^i (\mathring{d}_i + \hat{\epsilon}_i) + \sum_{j=1}^s v_\alpha^j (\mathring{e}_j + \check{\epsilon}_j)$$

where, by Lemma L.7, $\hat{\epsilon} \in \mathbb{R}^r, \check{\epsilon} \in \mathbb{R}^s$ are vectors that go to 0 almost surely with $n$. Then we apply the smoothness bound Eq. (76) to get, for each $\alpha \in J$

$$\begin{aligned}\left|\Psi_\alpha\left(\omega_\alpha; \sigma^2 d_\alpha\right) - \Psi_\alpha\left(\sum_{i=1}^r \mathring{d}_i x_\alpha^i + \sum_{j=1}^s \mathring{e}_j v_\alpha^j; \mathring{\sigma}^2\right)\right| &\leq R\left(1 + \min(\sigma^2 d_\alpha, \mathring{\sigma}^2)^{-1}\right) X_\alpha Y_\alpha \\ &\leq R\left(1 + \min(\sigma^2/2, \mathring{\sigma}^2)^{-1}\right) X_\alpha Y_\alpha\end{aligned}$$

using the definition of $J$ that $d_\alpha \geq 1/2, \forall \alpha \in J$. Here we have defined

$$\begin{aligned}X_\alpha &\stackrel{\text{def}}{=} |\omega_\alpha - \sum_{i=1}^r \mathring{d}_i x_\alpha^i - \sum_{j=1}^s \mathring{e}_j v_\alpha^j| + |\sigma^2 d_\alpha - \mathring{\sigma}^2| \\ &= \left|\sum_{i=1}^r x_\alpha^i \hat{\epsilon}_i + \sum_{j=1}^s v_\alpha^j \check{\epsilon}_j\right| + |\sigma^2 d_\alpha - \mathring{\sigma}^2| \\ Y_\alpha &\stackrel{\text{def}}{=} S_\alpha + |\omega_\alpha|^p + \left|\sum_{i=1}^r x_\alpha^i \hat{\epsilon}_i + \sum_{j=1}^s v_\alpha^j \check{\epsilon}_j\right|^p + \max(\sigma^2 d_\alpha, \mathring{\sigma}^2)^{p/2} + |\sigma^2 d_\alpha - \mathring{\sigma}^2|^{p/2}.\end{aligned}$$

Thus,

$$\begin{aligned}\textit{Eq. (74)} &= \frac{1}{n}\sum_{\alpha \in J}\left|\Psi_\alpha\left(\omega_\alpha; \sigma^2 d_\alpha\right) - \Psi_\alpha\left(\sum_{i=1}^r \mathring{d}_i x_\alpha^i + \sum_{j=1}^s \mathring{e}_j v_\alpha^j; \mathring{\sigma}^2\right)\right| \\ &\leq R\frac{1}{n}\left(1 + \min(\sigma^2/2, \mathring{\sigma}^2)^{-1}\right)\sum_{\alpha \in J} X_\alpha Y_\alpha \\ &\leq R\left(1 + \min(\sigma^2/2, \mathring{\sigma}^2)^{-1}\right)\sqrt{\frac{1}{n}\sum_{\alpha \in J} X_\alpha^2}\sqrt{\frac{1}{n}\sum_{\alpha \in J} Y_\alpha^2}.\end{aligned}$$

Since $\sigma \xrightarrow{\text{a.s.}} \mathring{\sigma}$ and we have assumed $\mathring{\sigma} > 0$ by Eq. ($\star$), we have $\left(1 + \min(\sigma^2/2, \mathring{\sigma}^2)^{-1}\right)$ is almost surely uniformly bounded in $n$.

Therefore, Eq. (74) can be shown to converge a.s. to 0 if we show

$$\sqrt{\frac{1}{n}\sum_{\alpha \in J} Y_\alpha^2} \quad \text{is a.s. uniformly bounded in } n \text{, and}$$

$$\sqrt{\frac{1}{n}\sum_{\alpha \in J} X_\alpha^2} \xrightarrow{\text{a.s.}} 0$$

We prove these two claims in Lemmas L.22 and L.23 below, which would finish our proof of $\mathsf{B} \xrightarrow{\text{a.s.}} 0$, and of our main theorem Theorem 2.10 as well.

**Lemma L.22.** $\sqrt{\frac{1}{n}\sum_{\alpha \in J} X_\alpha^2} \xrightarrow{\text{a.s.}} 0.$

*Proof.* Note that

$$\begin{aligned}X_\alpha &\leq \left|\sum_{i=1}^r x_\alpha^i \hat{\epsilon}_i + \sum_{j=1}^s v_\alpha^j \check{\epsilon}_j\right| + |\mathring{\sigma}^2 - \sigma^2| + |\sigma^2 - \sigma^2 d_\alpha| \\ &\stackrel{\text{def}}{=} P_\alpha + Q_\alpha + R_\alpha.\end{aligned}$$

Then by triangle inequality (in $\ell_2$-norm),

$$\sqrt{\frac{1}{n}\sum_{\alpha \in J} X_\alpha^2} \leq \sqrt{\frac{1}{n}\sum_{\alpha \in J} P_\alpha^2} + \sqrt{\frac{1}{n}\sum_{\alpha \in J} Q_\alpha^2} + \sqrt{\frac{1}{n}\sum_{\alpha \in J} R_\alpha^2}.$$

We now show that each term above converges a.s. to 0, which would finish the proof of Lemma L.22.

- Because $\hat{\epsilon} \xrightarrow{\text{a.s.}} 0$ and $\check{\epsilon} \xrightarrow{\text{a.s.}} 0$, we have, for some constant $C > 0$,

$$\begin{aligned}
\frac{1}{n}\sum_{\alpha\in J} P_\alpha^2 &\le C\frac{1}{n}\sum_{\alpha\in J}\left(\sum_{i=1}^r (x_\alpha^i\hat{\epsilon}_i)^2 + \sum_{j=1}^s (v_\alpha^j\check{\epsilon}_j)^2\right)\\
&\le C\max_{i,j}\{|\hat{\epsilon}_i|, |\check{\epsilon}_j|\} \times \frac{1}{n}\sum_{\alpha\in J}\left(\sum_{i=1}^r (x_\alpha^i)^2 + \sum_{j=1}^s (v_\alpha^j)^2\right)\\
&\le C\max_{i,j}\{|\hat{\epsilon}_i|, |\check{\epsilon}_j|\} \times \frac{1}{n}\sum_{\alpha\in [n]}\left(\sum_{i=1}^r (x_\alpha^i)^2 + \sum_{j=1}^s (v_\alpha^j)^2\right)\\
&\xrightarrow{\text{a.s.}} C \times 0 \times \mathcal{E} = 0
\end{aligned}$$

  where $\mathcal{E}$ is the Gaussian expectation that $\frac{1}{n}\sum_{\alpha\in[n]}\left(\sum_{i=1}^r (x_\alpha^i)^2 + \sum_{j=1}^s (v_\alpha^j)^2\right)$ converges a.s. to, by inductive hypothesis.

- The quantity $Q_\alpha$ actually doesn't depend on $\alpha$, so that

$$\sqrt{\frac{1}{n}\sum_{\alpha\in J} Q_\alpha^2} \le |\mathring{\sigma}^2 - \sigma^2| \xrightarrow{\text{a.s.}} 0$$

  by Lemma L.6.

- Notice $R_\alpha^2 = \sigma^4(1-d_\alpha)^2 \le \sigma^4(1-d_\alpha)$ because $1 - d_\alpha \in [0, 1/2]$ for $\alpha \in J$. Thus,

$$\frac{1}{n}\sum_{\alpha\in J} R_\alpha^2 \le \sigma^4\frac{1}{n}\sum_{\alpha\in J} 1 - d_\alpha \le \sigma^4\frac{1}{n}\sum_{\alpha\in[n]} 1 - d_\alpha = \sigma^4\frac{1}{n}\operatorname{rank}\mathbf{V}$$

  by the definition that $d_\alpha = (\mathbf{\Pi}_{\mathbf{V}}^{\perp})_{\alpha\alpha}$. But of course $\operatorname{rank}\mathbf{V} \le s$ is bounded relative to $n$. So this quantity goes to 0 almost surely) as desired.

$\square$

**Lemma L.23.** *$\sqrt{\frac{1}{n}\sum_{\alpha\in J} Y_\alpha^2}$ is a.s. uniformly bounded in $n$.*

*Proof.* We have

$$\begin{aligned}
\sqrt{\frac{1}{n}\sum_{\alpha\in J} Y_\alpha^2} &\le \sqrt{\frac{1}{n}\sum_{\alpha\in J} S_\alpha^2} + \sqrt{\frac{1}{n}\sum_{\alpha\in J} |\omega_\alpha|^{2p}} + \sqrt{\frac{1}{n}\sum_{\alpha\in J} \hat{X}_\alpha{}^2} + \sqrt{\frac{1}{n}\sum_{\alpha\in J} \max(\sigma^2 d_\alpha, \mathring{\sigma}^2)^p}\\
&\le \sqrt{\frac{1}{n}\sum_{\alpha\in [n]} S_\alpha^2} + \sqrt{\frac{1}{n}\sum_{\alpha\in [n]} |\omega_\alpha|^{2p}} + \sqrt{\frac{1}{n}\sum_{\alpha\in [n]} \hat{X}_\alpha{}^2} + \sqrt{\frac{1}{n}\sum_{\alpha\in [n]} \max(\sigma^2 d_\alpha, \mathring{\sigma}^2)^p}
\end{aligned}$$

where

$$\hat{X}_\alpha \stackrel{\text{def}}{=} \left|\sum_{i=1}^r x_\alpha^i\hat{\epsilon}_i + \sum_{j=1}^s v_\alpha^j\check{\epsilon}_j\right|^p + |\sigma^2 d_\alpha - \mathring{\sigma}^2|^{p/2}$$

We proceed to show that each of 4 summands above are individually a.s. uniformly bounded in $n$.

- $S_\alpha^2$ is a polynomially bounded function of $g_\alpha^1, \ldots, g_\alpha^{m-1}$, so that by Moments$(m-1)$,

$$\frac{1}{n}\sum_{\alpha\in[n]} S_\alpha^2 \xrightarrow{\text{a.s.}} C$$

  for some constant $C$, so it is also a.s. uniformly bounded in $n$.

- By Lemma L.20, we get
$$\frac{1}{n}\sum_{\alpha\in[n]}|\omega_\alpha|^{2p}$$
is a.s. uniformly bounded in $n$.

- Using the same reasoning as in the proof of Lemma L.22, one can easily show
$$\frac{1}{n}\sum_{\alpha\in[n]}\hat{X}_\alpha{}^2 \xrightarrow{\text{a.s.}} 0$$
so it is also a.s. uniformly bounded.

- Since $d_\alpha \le 1$, we have $\max(\sigma^2 d_\alpha, \mathring{\sigma}^2) \le \max(\sigma^2, \mathring{\sigma}^2)$, which is independent of $\alpha$. Therefore,
$$\frac{1}{n}\sum_{\alpha\in[n]}\max(\sigma^2 d_\alpha, \mathring{\sigma}^2)^p \le \frac{1}{n}\sum_{\alpha\in[n]}\max(\sigma^2, \mathring{\sigma}^2)^p = \max(\sigma^2, \mathring{\sigma}^2)^{p/2} \xrightarrow{\text{a.s.}} \mathring{\sigma}^p.$$
and it is also a.s. uniformly bounded in $n$.

□

Finally, we deliver the promised proof of Lemma L.21.

*Proof of Lemma L.21.* By Lemma K.4, $\Psi_\alpha$ is differentiable in $w$, and
$$\partial_w \Psi_\alpha(w;\tau^2) = \tau^{-1} \mathop{\mathbb{E}}_{z\sim\mathcal{N}(0,1)} z\psi(g_\alpha^1,\dots,g_\alpha^{m-1}, w+\tau z) \tag{77}$$
$$\partial_{\tau^2} \Psi_\alpha(w;\tau^2) = \frac{1}{2}\tau^{-2} \mathop{\mathbb{E}}_{z\sim\mathcal{N}(0,1)} (z^2-1)\psi(g_\alpha^1,\dots,g_\alpha^{m-1}, w+\tau z). \tag{78}$$

Recall that $|\psi(x)| \le C\|x\|_p^p + C$. We will silently introduce constants $C_1, C_2, \dots$ depending only on $p$, merging with old constants, typically via Lemma K.1 or by integrating out some integrands depending only on $p$. With $z \sim \mathcal{N}(0,1)$,
$$\begin{aligned}
|\partial_w \Psi_\alpha(w;\tau^2)| &\le \tau^{-1} \mathop{\mathbb{E}}_z |z||\psi(g_\alpha^1,\dots,g_\alpha^{m-1}, w+\tau z)| \\
&\le \tau^{-1} C \mathop{\mathbb{E}}_z |z| \left(1 + |g_\alpha^1|^p + \dots + |g_\alpha^{m-1}|^p + |w+\tau z|^p\right) \\
&\le \tau^{-1} C_1 \mathop{\mathbb{E}}_z |z| \left(1 + |g_\alpha^1|^p + \dots + |g_\alpha^{m-1}|^p + |w|^p + \tau^p|z|^p\right) \\
&\le \tau^{-1} C_2 \left(1 + |g_\alpha^1|^p + \dots + |g_\alpha^{m-1}|^p + |w|^p + \tau^p\right).
\end{aligned}$$

Similarly,
$$\begin{aligned}
|\partial_{\tau^2} \Psi_\alpha(w;\tau^2)| &\le \frac{1}{2}\tau^{-2} \mathop{\mathbb{E}}_z |z^2-1||\psi(g_\alpha^1,\dots,g_\alpha^{m-1}, w+\tau z)| \\
&\le \tau^{-2} C_3 \left(1 + |g_\alpha^1|^p + \dots + |g_\alpha^{m-1}|^p + |w|^p + \tau^p\right).
\end{aligned}$$

Therefore, for any $\Delta w \in \mathbb{R}, \Delta\tau^2 \in \mathbb{R}^{\ge 0}$, we have
$$\begin{aligned}
&\left|\Psi_\alpha(w+\Delta w;\tau^2+\Delta\tau^2) - \Psi_\alpha(w;\tau^2)\right| \\
&= \left|\int_0^1 \mathrm{d}t \left(\Delta w \cdot \partial_w \Psi_\alpha(w+\Delta w t;\tau^2+\Delta\tau^2 t) + \Delta\tau^2 \cdot \partial_{\tau^2}\Psi_\alpha(w+\Delta w t;\tau^2+\Delta\tau^2 t)\right)\right| \\
&\le \int_0^1 \mathrm{d}t \left(|\Delta w| \cdot |\partial_w \Psi_\alpha(w+\Delta w t;\tau^2+\Delta\tau^2 t)| + |\Delta\tau^2| \cdot |\partial_{\tau^2}\Psi_\alpha(w+\Delta w t;\tau^2+\Delta\tau^2 t)|\right) \\
&\le (C_2+C_3)(|\Delta w| + |\Delta\tau^2|) \\
&\quad \int_0^1 \mathrm{d}t((\tau^2+\Delta\tau^2 t)^{-1/2} + (\tau^2+\Delta\tau^2 t)^{-1}) \times \left(S_\alpha + |w+\Delta w t|^p + (\tau^2+\Delta\tau^2 t)^{p/2}\right)
\end{aligned}$$

where

$$S_\alpha \stackrel{\text{def}}{=} 1 + |g_\alpha^1|^p + \cdots + |g_\alpha^{m-1}|^p,$$

which is independent of $t$. Since $\Delta\tau^2 \geq 0$, $(\tau^2 + \Delta\tau^2 t)^{-1} \leq \tau^{-2}$, and we get

$$\begin{aligned}
&\left|\Psi_\alpha(w + \Delta w; \tau^2 + \Delta\tau^2) - \Psi_\alpha(w; \tau^2)\right| \\
&\leq C_4(|\Delta w| + \Delta\tau^2)(\tau^{-1} + \tau^{-2}) \int_0^1 \mathrm{d}t \left(S_\alpha + |w + \Delta w t|^p + (\tau^2 + \Delta\tau^2 t)^{p/2}\right) \\
&\leq C_5(|\Delta w| + \Delta\tau^2)(\tau^{-1} + \tau^{-2}) \int_0^1 \mathrm{d}t \left(S_\alpha + |w|^p + |\Delta w|^p t^p + \tau^p + (\Delta\tau^2)^{p/2} t^{p/2}\right) \\
&\leq C_6(|\Delta w| + \Delta\tau^2)(\tau^{-1} + \tau^{-2}) \left(S_\alpha + |w|^p + |\Delta w|^p + \tau^p + (\Delta\tau^2)^{p/2}\right)
\end{aligned}$$

where in the end we have integrated out $t^p$ and $t^{p/2}$. We finally apply the simplification $\tau^{-1} \leq \frac{1}{2} + \frac{1}{2}\tau^{-2}$ by AM-GM to get the desired Eq. (76). $\square$

# M Proof of NETSOR$\top^+$ Master Theorem Assuming Rank Stability

In this section we describe how to augment the proof of Theorem 2.10 given in Appendix L to yield the proof of Theorem E.11. This is very similar to the proof of the NETSOR$^+$ Master Theorem in Yang [50]. The key points to note here are 1) the rank stability assumption Assumption E.7 used in Theorem E.11, and 2) an additional term in Eq. (69) due to fluctuations in the parameter $\Theta$.

## M.1 Rank Stability

By Remark E.13, we see that rank stability assumption is necessary for the parameter-controlled NETSOR$\top^+$ Master Theorem. In the proof of the NETSOR$\top$ Master Theorem (Appendix L), we had to intricately weave together an induction on rank stability (more generally, CoreSet) and an induction on moment convergence (Moments). However, here, to show Theorem E.11, we just need 1) to induct on Moments and 2) to invoke Assumption E.7 whenever we need to use Lemma L.11, which is when we need to show that pseudo-inverse commutes with almost surely limit, such as in Proposition L.5, and when we need to ensure either $\sigma$ is almost surely 0 or is almost surely positive, as in Appendix L.7.3.

## M.2 Fluctuation of the Parameters

As in Appendix L, we will induct on the G-vars $g^1, \ldots, g^m$ in the program to show that

1. For any random vector $\Theta \in \mathbb{R}^l$ that converges almost surely to a deterministic vector $\mathring{\Theta}$ as $n \to \infty$, and for any $\psi(-;-) : \mathbb{R}^m \times \mathbb{R}^l \to \mathbb{R}$ parameter-controlled at $\mathring{\Theta}$,
$$\frac{1}{n}\sum_{\alpha=1}^n \psi(g_\alpha^1, \ldots, g_\alpha^m; \Theta) \xrightarrow{\text{a.s.}} \mathbb{E}\,\psi(Z^{g^1}, \ldots, Z^{g^m}; \mathring{\Theta}).$$
for any G-vars $g^1, \ldots, g^m$, where $Z^{g^i}$ are as defined in Definition E.3.
2. Each scalar $\theta$ that is a deterministic function of $g^1, \ldots, g^m$ converges a.s. to $\mathring{\theta}$.

The latter trivially follows from former, so we will focus on proving the former in the inductive step.

When we have parameters in nonlinearities, Eq. (69) needs to be modified to contain an additional term $\mathsf{D}$:

$$\left|\frac{1}{n}\sum_{\alpha=1}^n \psi(g_\alpha^1, \ldots, g_\alpha^m; \Theta) - \mathbb{E}\,\psi(Z^{g^1}, \ldots, Z^{g^m}; \mathring{\Theta})\right| \leq \mathsf{D} + \mathsf{A} + \mathsf{B} + \mathsf{C}$$

where

$$\mathsf{D} \stackrel{\text{def}}{=} \left|\frac{1}{n}\sum_{\alpha=1}^n \psi(g_\alpha^1, \ldots, g_\alpha^m; \Theta) - \psi(g_\alpha^1, \ldots, g_\alpha^m; \mathring{\Theta})\right|$$

and $\mathsf{A}, \mathsf{B}, \mathsf{C}$ are as in Eq. (69) but replacing $\psi(-)$ there with $\psi(-; \mathring{\Theta})$. Because $\psi(-;-)$ is parameter-controlled at $\mathring{\Theta}$ by assumption, $\psi(-; \mathring{\Theta})$ is controlled, and $\mathsf{A}, \mathsf{B}, \mathsf{C} \xrightarrow{\text{a.s.}} 0$ with the same arguments as before (except using rank stability assumption Assumption E.7 where appropriate, instead of CoreSet).

Now, by the other property of parameter-control, we have

$$\begin{aligned} \mathsf{D} \leq \frac{1}{n} \sum_{\alpha=1}^{n} \left| \psi(g_\alpha^1, \ldots, g_\alpha^m; \Theta) - \psi(g_\alpha^1, \ldots, g_\alpha^m; \mathring{\Theta}) \right| &\leq \frac{1}{n} \sum_{\alpha=1}^{n} f(\Theta) \bar{\psi}(g_\alpha^1, \ldots, g_\alpha^m) \\ &= f(\Theta) \frac{1}{n} \sum_{\alpha=1}^{n} \bar{\psi}(g_\alpha^1, \ldots, g_\alpha^m) \end{aligned}$$

for some controlled $\bar{\psi} : \mathbb{R}^m \to \mathbb{R}$ and some $f : \mathbb{R}^l \to \mathbb{R}^{\geq 0} \cup \{\infty\}$ that is continuous at $\mathring{\Theta}$ and has $f(\mathring{\Theta}) = 0$ (where $\bar{\psi}$ and $f$ can both depend on $\mathring{\Theta}$). Since $\Theta \xrightarrow{\text{a.s.}} \mathring{\Theta}$ by induction hypothesis, we have $f(\Theta) \xrightarrow{\text{a.s.}} 0$. In addition, by Moments, $\frac{1}{n} \sum_{\alpha=1}^{n} \bar{\psi}(g_\alpha^1, \ldots, g_\alpha^m)$ converges a.s. as well to a finite constant. Therefore,

$$\mathsf{D} \xrightarrow{\text{a.s.}} 0$$

as desired.

### M.3 Summary

The proof of Theorem E.11 would proceed as follows: We induct on Moments with the same setup as Appendix L.4, except using Assumption E.7 for Proposition L.5. Then we prove the inductive step for Moments as in Appendix L.7. We modify Eq. (69) to add a term $\mathsf{D}$ as in Appendix M.2, which goes to 0 a.s. as argued there. The same arguments for $\mathsf{A}, \mathsf{B}, \mathsf{C} \xrightarrow{\text{a.s.}} 0$, exhibited in Appendix L.7 still hold, except that in the proof of $\mathsf{B} \xrightarrow{\text{a.s.}} 0$, we apply Assumption E.7 (instead of Lemma L.11) to allow us to assume $\mathring{\sigma} > 0$ and $\sigma > 0$ almost surely.

## N Proof of NETSOR$\top^+$ Master Theorem without Assuming Rank Stability

In this section, we prove Theorem E.15.

**Definition N.1.** We say a vector $v \in \mathbb{R}^n$ has *vanishing moments* if $\frac{1}{n} \sum_{\alpha=1}^{n} v_\alpha^{2k} \xrightarrow{\text{a.s.}} 0$ for all integer $k > 0$ as $n \to \infty$. We say it has *bounded moments* if there are finite $C_k \in \mathbb{R}$ such that $\frac{1}{n} \sum_{\alpha=1}^{n} v_\alpha^{2k} \xrightarrow{\text{a.s.}} C_k$ for all $k$.

Assume each `Nonlin`$^+$ instruction only takes G-vars instead of any vector in the program.

**Proof Plan** We will inductively rewrite the program and maintain a 1) core set $\mathcal{M}$ of G-vars whose elements are orthogonal as vectors and have non-vanishing moments and 2) a set $\overline{\mathcal{M}}$ of G-vars with vanishing moments. In particular, the vectors in $\overline{\mathcal{M}}$ will take the form $Av$ for some matrix $A$ and some $v$ with vanishing moments. Any vector in the original program will always be (mathematically) a linear combination of vectors in $\mathcal{M}$ and $\overline{\mathcal{M}}$. Then Moments largely reduces to that of $\mathcal{M}$, which has rank stability.

Every non-initial vector $g$ of $\mathcal{M}$ is created by `MatMul`, say $g = Ah$ for some matrix $A$ and vector $h$. For any matrix $A$, define $\mathcal{M}_A^*$ to be the collection of all such $h$.

### N.1 Induction Hypotheses

Let $g^1, \ldots, g^M$ be all of the G-vars in the program, including initial vectors, in order of appearance. We inductively rewrite the program and expand $\mathcal{M}$ and $\overline{\mathcal{M}}$, maintaining the following induction hypothesis for each $m \in [M]$.

**IH**$(m)$ The following hold simultaneously

**Rewrite**$(m)$ $g^1, \ldots, g^m$ are mathematically linear combinations of vectors in $\mathcal{M}$ and $\overline{\mathcal{M}}$ whose coefficients are definable using `Moment` applied to $\mathcal{M}$ and $\overline{\mathcal{M}}$.

**Moments**$(m)$ Let $z^1, \ldots, z^k$ be all vectors in $\mathcal{M}$ and $\bar{z}^1, \ldots, \bar{z}^{\bar{k}}$ be all vectors in $\overline{\mathcal{M}}$. For any pseudo-Lipschitz $\psi : \mathbb{R}^{k+\bar{k}} \to \mathbb{R}$,

$$\frac{1}{n}\sum_{\alpha=1}^{n} \psi(z^1_\alpha, \ldots, z^k_\alpha, \bar{z}^1_\alpha, \ldots, \bar{z}^{\bar{k}}_\alpha) \xrightarrow{\text{a.s.}} \mathbb{E}\,\psi(Z^{z^1}, \ldots, Z^{z^k}, 0, \ldots, 0).$$

**Orthogonal**$(m)$ For all $n$, for any matrix $A$, $\mathcal{M}^*_A$ forms an orthogonal system of vectors where each vector has $\ell_2$ norm $\sqrt{n}$

**VanSet**$(m)$ Every vector of $\overline{\mathcal{M}}$ has vanishing moments

By induction, we will have Rewrite$(M)$ and Moments$(M)$, which together with Lemma N.3 imply Theorem E.15.

### N.2 Base Case

WLOG, we may assume the initial vectors have identity covariance by applying Gram-Schmidt. These vectors are added to $\mathcal{M}$. Currently, $\overline{\mathcal{M}}$ is empty. Moments is then automatic by the Strong Law of Large Numbers; the other clauses of the induction hypothesis are obvious.

### N.3 Induction

Assume IH$(m-1)$. We seek to show IH$(m)$.

Suppose $g = g^m = Ah$ for some vector $h$. By Rewrite$(m-1)$, we have

$$h = \phi(z^1, \ldots, z^k, \bar{z}^1, \ldots, \bar{z}^{\bar{k}}; \theta_1, \ldots, \theta_l)$$

where 1) $z^1, \ldots, z^k$ are all of the vectors in $\mathcal{M}$, 2) $\bar{z}^1, \ldots, \bar{z}^{\bar{k}}$ are all of the vectors in $\overline{\mathcal{M}}$, 3) $\theta_1, \ldots, \theta_l$ are scalars created by `Moment` from $z^1, \ldots, z^k, \bar{z}^1, \ldots, \bar{z}^{\bar{k}}$ and so converge a.s. to $\mathring{\theta}_1, \ldots, \mathring{\theta}_l$ by Moments$(m-1)$, and 4) $\phi$ is pseudo-Lipschitz jointly in all of them.

Let

$$h^0 \stackrel{\text{def}}{=} \phi(z^1, \ldots, z^k, 0, \ldots, 0; \theta_1, \ldots, \theta_l) \quad \text{and} \quad \Delta h \stackrel{\text{def}}{=} h - h^0.$$

(Note both $h^0$ and $\Delta h$ are expressible by $\texttt{Nonlin}^+$ in terms of $z^1, \ldots, z^k, \bar{z}^1, \ldots, \bar{z}^{\bar{k}}; \theta_1, \ldots, \theta_l$). Because $\bar{z}^1, \ldots, \bar{z}^{\bar{k}}$ have vanishing moments by induction hypothesis and $\phi$ is pseudo-Lipschitz, $\Delta h$ has vanishing moments as well by Lemma N.3. We insert $A\Delta h$ into $\overline{\mathcal{M}}$; we will prove that $A\Delta h$ has vanishing moments below in Lemma N.2.

Now define $\hat{h}^0 \stackrel{\text{def}}{=} h^0 - \Pi h^0$ where $\Pi$ is the projection onto the linear span of $\mathcal{M}^*_A$. $\Pi h^0$ can be written as a $\texttt{Nonlin}^+$ like so: $\Pi h^0 = \sum_{v \in \mathcal{M}^*_A} \frac{v^\top h^0}{n} v$, where we used the induction hypothesis that $\|v\|_2 = \sqrt{n}$ and $\mathcal{M}^*_A$ is orthogonal. Then $\hat{h}^0$ can also be written as a $\texttt{Nonlin}^+$.

We proceed by casework on whether $Z^{\hat{h}^0} = 0$.

1) Suppose $Z^{\hat{h}^0} = 0$. Then $A\hat{h}^0$ has vanishing moments by the same Gaussian conditioning technique as in Appendix L, which is possible because we only need to condition on $z^1, \ldots, z^k$, which are all of the G-vars that $\hat{h}^0$ depends on, and Orthogonal$(m-1)$ implies they have rank stability. Then we add $A\hat{h}^0$ to $\overline{\mathcal{M}}$. Because $g = Ah = A(h^0 + \Delta h) = A(\hat{h}^0 + \Pi h^0 + \Delta h)$, we can rewrite all instances of $g$ in the program as the sum of $A\hat{h}^0$, $A\Delta h$ (in $\overline{\mathcal{M}}$), and $A\Pi h^0 = \sum_{v \in \mathcal{M}^*_A} \frac{v^\top h^0}{n} Av$ (a linear combination of $\mathcal{M}$).

2) Suppose $Z^{\hat{h}^0} \neq 0$. Define $h^1 \stackrel{\text{def}}{=} h^0 / \frac{\|\hat{h}^0\|^2}{n}$ via $\texttt{Nonlin}^+$ (which is valid since $\frac{\|\hat{h}^0\|^2}{n} \xrightarrow{\text{a.s.}} \mathbb{E}(Z^{\hat{h}^0})^2 \neq 0$ by induction hypothesis). We add $Ah^1$ to $\mathcal{M}$. Because $g = Ah = A(h^0 + \Delta h) = A(\frac{\|\hat{h}^0\|^2}{n} h^1 + \Pi h^0 + \Delta h)$, we can rewrite all instances of $g$ in the program as the sum of $A\Delta h$ (in $\overline{\mathcal{M}}$), $\frac{\|\hat{h}^0\|^2}{n} Ah^1$, and $A\Pi h^0 = \sum_{v \in \mathcal{M}^*_A} \frac{v^\top h^0}{n} Av$ (linear combinations of $\mathcal{M}$).

With the above updates to $\mathcal{M}$ and/or $\overline{\mathcal{M}}$, it's clear that Rewrite$(m)$, Orthogonal$(m)$ and VanSet$(m)$ are true. So it remains to prove Moments$(m)$.

**Moments**$(m)$ In the case 1) above (where we did not expand $\mathcal{M}$), Moments$(m)$ follows straightforwardly from Moments$(m-1)$ and Lemma N.3.

In the case 2) above (where we added $Ah^1$ to $\mathcal{M}$), we need to show

$$\frac{1}{n}\sum_{\alpha=1}^{n}\psi(Ah^1_\alpha, z^1_\alpha, \dots, z^k_\alpha, \bar{z}^1_\alpha, \dots, \bar{z}^{\bar{k}}_\alpha) \xrightarrow{\text{a.s.}} \mathbb{E}\,\psi(Z^{Ah^1_\alpha}, Z^{z^1}, \dots, Z^{z^k}, 0, \dots, 0)$$

for any pseudo-Lipschitz $\psi$. Since $h^1$ is expressible as a `Nonlin`$^+$ of purely $z^1, \dots, z^k$ and some parameters, we can just condition on $z^1_\alpha, \dots, z^k_\alpha$ and proceed as in Appendix L, after replacing $\bar{z}^1_\alpha, \dots, \bar{z}^{\bar{k}}_\alpha$ by zeros by Lemma N.3. (Note this argument wouldn't have worked if we naively put $g = Ah$ in place of $Ah^1$ because $h$ depends on $\bar{z}^1, \dots, \bar{z}^{\bar{k}}$).

This finishes the induction, barring for the promised Lemma N.2 below.

**Lemma N.2.** *$A\Delta h$ has vanishing moments.*

It's clear that $A\Delta h$ at least has vanishing second moment because $A$ has a uniformly (in $n$) bounded operator norm from standard matrix concentration bounds. We need to work a bit harder to get all moments to vanish.

*Proof.* We condition on $z^1, \dots, z^k, \bar{z}^1, \dots, \bar{z}^{\bar{k}}$. Suppose

$$u^j = A^\top v^j, \quad j = 1, \dots, s, \quad \text{and} \quad \bar{u}^j = A^\top \bar{v}^j, \quad j = 1, \dots, \bar{s}$$

are respectively the elements of $\mathcal{M}$ and $\overline{\mathcal{M}}$ that are images of MatMul with $A^\top$. As in the ($\hat{Z}$, $\dot{Z}$)-decomposition in the $n \to \infty$ limit, we can write $A\Delta h$, for finite $n$, as a sum of 1) a Gaussian vector with vanishing 2nd moment and 2) a linear combination of $\{v^j\}_{j=1}^{s}$ and $\{\bar{v}^j\}_{j=1}^{\bar{s}}$, where the coefficient of $v^j$ (resp. $\bar{v}^j$) converges to $\sigma_A^2\,\mathbb{E}\,\partial Z^{\Delta h}/\partial Z^{v^j} = 0$ (resp. $\sigma_A^2\,\mathbb{E}\,\partial Z^{\Delta h}/\partial Z^{\bar{v}^j} < \infty$). The former has vanishing moments of all order by Gaussianity. Since $v^j$ has bounded moments and $\bar{v}^j$ has vanishing moments, the latter also has vanishing moments. ☐

**Lemma N.3.** *Let $z^1, \dots, z^k$ have bounded moments, $\bar{z}^1, \dots, \bar{z}^{\bar{k}}$ have vanishing moments, and $\theta_1, \dots, \theta_l \in \mathbb{R}$ converge almost surely to deterministic $\mathring{\theta}_1, \dots, \mathring{\theta}_l \in \mathbb{R}$. Then for any pseudo-Lipschitz $\psi$ and $t > 0$,*

$$\frac{1}{n}\sum_{\alpha=1}^{n}|\psi(z^1_\alpha, \dots, z^k_\alpha, \bar{z}^1_\alpha, \dots, \bar{z}^{\bar{k}}_\alpha; \theta_1, \dots \theta_l) - \psi(z^1_\alpha, \dots, z^k_\alpha, 0, \dots, 0; \mathring{\theta}_1, \dots, \mathring{\theta}_l)|^t \xrightarrow{\text{a.s.}} 0.$$

*Proof.* Holder's inequality. ☐