Factoring Matrices into Linear Neural Networks

Sagnik Bhattacharya
Jonathan Shewchuk

Electrical Engineering and Computer Sciences
University of California, Berkeley

Technical Report No. UCB/EECS-2022-92
http://www2.eecs.berkeley.edu/Pubs/TechRpts/2022/EECS-2022-92.html

May 13, 2022

Acknowledgement

I would like to thank my advisor Professor Jonathan Shewchuk who helped me push through the numerous sticking points in the research process, Dr. Marc Khoury who opened the door to research at UC Berkeley, my parents who were always there to support me, my mentors who helped me with various aspects of my undergraduate, graduate, and post-graduate life, and my friends who made it all memorable.

Factoring Matrices into Linear Neural Networks

by Sagnik Bhattacharya

Research Project

Submitted to the Department of Electrical Engineering and Computer Sciences, University of California at Berkeley, in partial satisfaction of the requirements for the degree of **Master of Science, Plan II**.

Approval for the Report and Comprehensive Examination:

Committee:

Professor Jonathan R. Shewchuk
Research Advisor

12 May 2022
(Date)

****

James O'Brien
Second Reader

13 May 2022
(Date)

Acknowledgements

I would like to thank my advisor Professor Jonathan Shewchuk who helped me push through the numerous sticking points in the research process, Dr. Marc Khoury who opened the door to research at UC Berkeley, my parents who were always there to support me, my mentors who helped me with various aspects of my undergraduate, graduate, and post-graduate life, and my friends who made it all memorable.

Factoring a Matrix into Linear Neural Networks

Sagnik Bhattacharya

Master of Science in Electrical Engineering and Computer Sciences

University of California, Berkeley

Professor Jonathan R. Shewchuk, Chair

Abstract

We characterize the topology and geometry of the set of all weight vectors for which a linear neural network computes the same linear transformation W. This set of weight assignments is called the *fiber* of W, and it is embedded in a Euclidean *weight space* of all possible weight vectors. The fiber is an algebraic variety with singular points, hence it is not a manifold. We show a way to *stratify* the fiber—that is, to partition the algebraic variety into a finite set of manifolds of varying dimensions called *strata*. We derive the dimensions of these strata and the relationships by which they adjoin each other. (Although they are disjoint, some strata lie in the closures of other, higher-dimensional strata.) Each stratum is smoothly embedded in weight space, so it has a well-defined tangent space (which is a subspace of weight space) at every point. We show how to determine the subspace tangent to a specified stratum at a specified point on the stratum, and we construct an elegant basis for that subspace.

To help achieve these goals, we first derive a *Fundamental Theorem of Linear Neural Networks*, analogous to Gilbert Strang's *Fundamental Theorem of Linear Algebra*. We show how to decompose each layer of a linear neural network into a set of subspaces that show how information flows through the neural network—in particular, tracing which information is annihilated at which layers of the network, and identifying subspaces that carry no information but might become available to carry information as training modifies the network weights. We summarize properties of these information flows in "basis flow diagrams" that reveal a rich and occasionally surprising structure. Each stratum of the fiber represents a different pattern by which information flows (or fails to flow) through the neural network.

We use this knowledge to find transformations in weight space called *moves* that allow us to modify the neural network's weights without changing the linear transformation that the network computes. Some moves stay on the same stratum, and some move from one stratum to another stratum of the fiber. In this way, we can visit different weight assignments for which the neural network computes the same transformation. These moves help us to construct a useful basis for the weight space and a useful basis for each space tangent to a stratum.

1 Introduction

In its simplest form, a *linear neural network* is a sequence of matrices whose product is a matrix. The first matrix linearly transforms an input vector; each subsequent matrix linearly transforms the vector produced by the previous matrix; and the composition of those transformations is also a linear transformation, represented by the product of the matrices. But this definition is incomplete: the term of art *neural network* typically also entails software that computes the sequence of vectors, an optimization algorithm that *trains* the network by choosing good weights, and more. So here is a computational definition: a *linear neural network* is a neural network in which there are no activation functions. (If you prefer, the activation function at the output of each unit is just the identity function.) Each layer of connections (edges) in the network is represented by a matrix—for layer k of edges, we call it W_k. Given an input vector x, the network computes the linear transformation

$$y = Wx, \qquad \text{where } W = \mu(W_L, W_{L-1}, \ldots, W_2, W_1) = W_L W_{L-1} \cdots W_2 W_1.$$

The matrices are numbered in the order they are applied in computation. For brevity, we omit added terms, which do not appreciably affect our results.

In this paper, we study $\mu^{-1}(W)$, the set of all factorizations of a matrix W into a product of matrices of specified sizes. This set is infinite and it is an *algebraic variety*—the set of all solutions of a system of polynomial equations. Trager, Kohn, and Bruna [16] call $\mu^{-1}(W)$ the *fiber* of W. Said differently, we wish to study the set of all choices of linear neural network weights such that the network computes the linear transformation W. The fiber has a complicated topology and geometry: it is a union of manifolds of varying dimensions. Understanding the fiber has applications in understanding gradient descent algorithms for training neural networks—but it is also a beautiful mathematical problem in its own right.

We also study how to move along the fiber. This study is motivated by our belief that movements along a fiber can sometimes help a neural network to improve its training speed, and that performing such movements is practical even when computing W explicitly is not.

Let us give a simple example of a fiber. Suppose every matrix is square and W is invertible; then every factor matrix W_i must also be invertible. The set of all real, invertible $d \times d$ matrices is called the *general linear group* $\mathrm{GL}(d, \mathbb{R})$, which is a d^2-dimensional manifold embedded in $\mathbb{R}^{d \times d}$. $\mathrm{GL}(d, \mathbb{R})$ has two connected components: one for matrices with positive determinants and one for negative determinants. To factor W, we can choose each matrix W_i to be an arbitrary member of $\mathrm{GL}(d, \mathbb{R})$ except for one matrix that is uniquely determined by the other choices. The fiber is

$$\mu^{-1}(W) = \{(W_L, W_{L-1}, \ldots, W_2, W_2^{-1} W_3^{-1} \cdots W_L^{-1} W) : W_L, W_{L-1}, \ldots, W_2 \in \mathrm{GL}(d, \mathbb{R})\}.$$

This fiber is a smooth, $(d^2(L-1))$-dimensional manifold with topology $\mathrm{GL}(d, \mathbb{R}) \times \mathrm{GL}(d, \mathbb{R}) \times \ldots \mathrm{GL}(d, \mathbb{R})$ (with $L-1$ factors) and hence 2^{L-1} connected components (reflecting the signs of the determinants of the factor matrices). Figure 1 graphs the fiber $\mu^{-1}([1])$ when we factor the matrix [1] into three 1×1 matrices. Although this graph lacks the complexities of larger matrices, we see a graceful 2-dimensional manifold with four components, as advertised, and we gain an inkling of what the general case might look like.

Unfortunately, if the network has a matrix that isn't square or if W does not have full rank, the fiber is usually no longer a manifold. But it can be partitioned into smooth manifolds of different dimensions, called *strata*, as illustrated in Figure 2, which charts the solutions of $[\theta_2][\theta_1 \quad \theta_1'] = [0 \quad 0]$. Each stratum represents a different pattern by which information flows (or fails to flow) through the neural network. As a side effect of understanding these strata, we will expose some fundamental properties of linear algebra that clarify how subspaces are mapped from layer to layer and which subspaces ultimately vanish into the nullspaces of

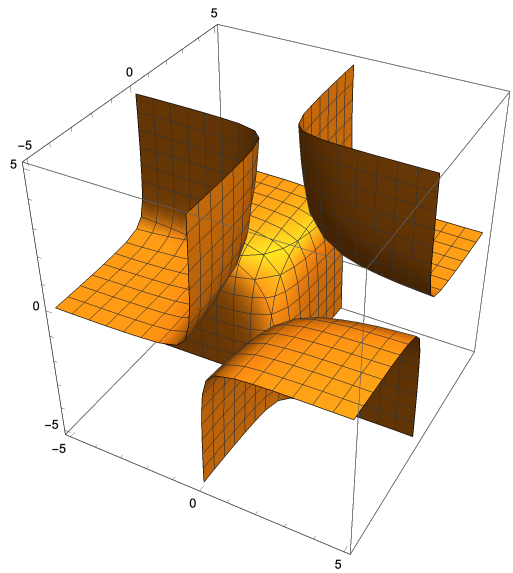


Figure 1: The fiber $\mu^{-1}([1])$ for the network $W_3W_2W_1 = [\theta_3][\theta_2][\theta_1] = [1] = W$.

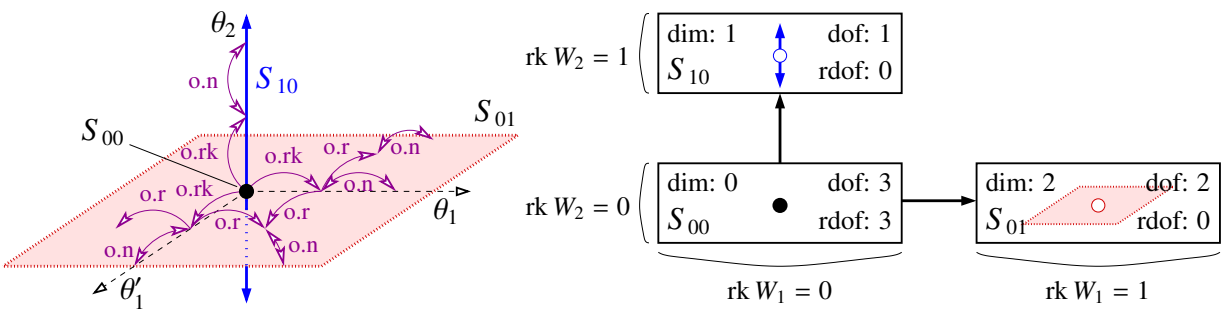


Figure 2: At left is the fiber $\mu^{-1}([0 \quad 0])$ for the network $W_2W_1 = [\theta_2][\theta_1 \quad \theta_1'] = [0 \quad 0] = W$, partitioned into three strata: S_{00} is the origin; S_{10} is the θ_2-axis with the origin removed; and S_{01} is the plane spanned by the θ_1- and θ_1'-axes with the origin removed. The purple arcs show examples of three types of moves: type o.rk moves modify one of the two matrices (W_2 or W_1) and increase its rank, thereby moving to a higher-dimensional stratum; type o.r moves modify one matrix (here, W_1) without changing its rank nor leaving the current stratum, but do change its rowspace; and type o.n moves modify one matrix without changing its rank or rowspace (or columnspace). At right, the strata are arranged in a dag, which is organized as a two-dimensional table indexed by the ranks of W_1 and W_2. Each dag vertex specifies the dimension of the stratum (dim), the number of degrees of freedom of motion on the fiber (dof), and the number of rank-increasing degrees of freedom (rdof) that generate o.rk moves off the stratum. Always, dof = dim + rdof. A directed edge from one stratum to another implies that the former lies in the closure of the latter, so an infinitesimal move can take you from the former into the latter.

which matrices. This hidden structure has a strong influence on training, though W does not reveal it. (In Figure 2, the worst place to start training is the origin.)

To understand how the strata are connected to themselves and each other, we study a set of operators we call *moves* (purple arcs in Figure 2) that map one network factorization of W to another along carefully chosen basis directions in weight space. (The basis is different for each point in weight space, which is why the arcs in Figure 2 don't look like they're all from the same basis.) These moves have both theoretical and practical

motivations. The theoretical motivation is that they provide a great deal of intuition about the geometry and topology of the fiber of a matrix W. The practical motivation is that although two different neural networks might compute the same transformation, one might be much more amenable to training than the other. It is well known that during training, a neural network can fall near a critical point in the cost function that slows down network training but is "spurious" in the sense that it is not related to the transformation being learned; it is merely a side effect of how that transformation happens to be encoded in layers. Researchers studying these phenomena include Trager, Kohn, and Bruna [16]. Spurious critical points appear to be one of the reasons that deep neural networks typically learn more slowly than shallow ones. Our original motivation for this paper is that we want to find ways to move away from spurious critical points, thereby speeding up learning, without changing the function that a network has learned. (This paper doesn't solve that problem, but it's a first step along the way.)

Linear neural networks compute only linear transformations; they are far less powerful than networks with nonlinear activation functions such as rectified linear units (ReLUs, also known as ramp functions) and sigmoid functions (also known as logistic functions). Yet linear networks have become a popular object of study [3, 9, 11, 17]. Why? We cannot fully understand the training of ReLU-based networks—or probably any neural networks—if we do not understand linear networks [7]. Training a linear neural network with a gradient descent method is a nonlinear process [13], exhibiting surprising phenomena like implicit acceleration of training [1] and implicit regularization [2, 4, 6]. Similar results about implicit regularization due to the alignment of layer weights during training by gradient descent were found by Ji and Telgarsky [8] and generalized by Radhakrishnan et al. [12]. Trager, Kohn, and Bruna [16] show that the map μ and the fiber $\mu^{-1}(W)$ play a crucial role in characterizing cost functions of linear neural networks and understanding critical points in their cost functions. Some results about linear neural networks, especially those about the function represented by the network, generalize to ReLU networks with minor caveats. For example, Li and Sompolinsky [10] propose a theory similar to statistical mechanics to study the input-output behavior of linear neural networks; empirically their theory appears to hold for large classes of ReLU networks.

2 Notation

Let L be the number of matrices—that is, the number of layers of edges (connections) in the network. Alternating with the edge layers are $L + 1$ layers of units, numbered from 0 to L, in which layer j has d_j real-valued units that represent a vector in $\mathbb{R}^{d_j}$. Layer 0 is the *input layer*, layer L is the *output layer*, and between them are $L - 1$ *hidden layers*. The layers of edges are numbered from 1 to L, and the edge weights in edge layer j are represented by a real-valued $d_j \times d_{j-1}$ matrix W_j.

We collect all the neural network's weights in a *weight vector* $\theta = (W_L, W_{L-1}, \ldots, W_1) \in \mathbb{R}^{d_\theta}$, where $d_\theta = d_L d_{L-1} + d_{L-1} d_{L-2} + \ldots + d_1 d_0$ is the number of real-valued weights in the network (i.e., the number of connections). Recall the function $\mu(W_L, W_{L-1}, \ldots, W_2, W_1) = W_L W_{L-1} \cdots W_2 W_1$; we can abbreviate it to $\mu(\theta)$. Given a fixed weight vector θ, our linear neural network takes an *input vector* $x \in \mathbb{R}^{d_0}$ and returns an *output vector* $y = W_L W_{L-1} \cdots W_2 W_1 x$, with $y \in \mathbb{R}^{d_L}$. Hence, the network implicitly computes a linear transformation specified by the $d_L \times d_0$ matrix $W = \mu(\theta)$, yielding $y = Wx$.

The map μ is not bijective, so we define its preimage to be a set. Let $\mu^{-1}(W) = \{\theta : \mu(\theta) = W\}$ be the set of all factorizations of W for some fixed $d_L, d_{L-1}, \ldots, d_0$. We call $\mu^{-1}(W)$ the *fiber* of W, and we will treat it as a geometric object embedded in the space $\mathbb{R}^{d_\theta}$. Note that $\mu^{-1}(W)$ is empty if and only if $\operatorname{rk} W > \min_{1 \le j \le L-1} d_j$.

Given $\theta \in \mathbb{R}^{d_\theta}$, its *subsequence matrices* are all the matrices of the form $W_{k \sim i} = W_k W_{k-1} \cdots W_{i+1}$. The notation $W_{k \sim i}$ indicates that this matrix transforms a vector at unit layer i to produce a vector at unit layer

k. Note that $W = W_{L\sim 0}$ and $W_j = W_{j\sim j-1}$. We call each W_j a *factor matrix*. We use the convention that $W_{k\sim k} = I_{d_k \times d_k}$, the $d_k \times d_k$ identity matrix.

The *rank list* for a weight vector $\theta \in \mathbb{R}^{d_\theta}$ is a sequence that lists the rank of every subsequence matrix $W_{k\sim i}$ such that $L \geq k \geq i \geq 0$. The list includes the unit layer sizes $\operatorname{rk} W_{k\sim k} = d_k$. For example, for a network with $L = 3$ layers of edges, the rank list is $\langle d_4, d_3, d_2, d_1, \operatorname{rk} W_3, \operatorname{rk} W_2, \operatorname{rk} W_1, \operatorname{rk} W_3W_2, \operatorname{rk} W_2W_1, \operatorname{rk} W\rangle$. No two strata have the same rank list; the rank list plays a major role as the index that labels each stratum. We sometimes omit the ranks that are invariant for a specific fiber: the d_j's and $\operatorname{rk} W$. For example, in Figure 2, the subscripts of S are $\operatorname{rk} W_2$ and $\operatorname{rk} W_1$, as only these ranks vary.

3 A Foretaste of our Results: Two Matrices

The fiber $\mu^{-1}(W)$ is an algebraic variety (again, the set of all solutions of a system of polynomial equations). In general, the variety is not a manifold. Its local dimension can vary, and it can have points where it branches (like S_{00} in Figure 2) or is otherwise weirdly connected to itself. We address this complication by partitioning the fiber into a set of manifolds called *strata*.[1] The strata are pairwise disjoint, and the fiber is the union of these strata. This partition is called a *stratification* of the variety.

Figures 2 and 3 depict stratifications of two fibers. Figure 2 graphs the variety of solutions to $W_2W_1 = [\theta_2][\theta_1 \quad \theta_1'] = [0 \quad 0] = W$, illustrating that a fiber may have a mix of dimensionalities. There are two ways to achieve $W_2W_1 = [0 \quad 0]$: we can set $W_2 = [0]$ or we can set $W_1 = [0 \quad 0]$. The former solutions lie on the pink plane in Figure 2, and the latter solutions lie on the blue line. In our stratification, we have chosen to partition the fiber into three parts, with the origin being a 0-dimensional stratum, labeled S_{00}. This is motivated by the fact that S_{00} is the sole point from which we have three degrees of freedom of motion on the fiber: two degrees of freedom on the plane, and one on the line. Observe that two of these degrees of freedom can be combined, and one cannot: if we move along the line, we can change only the θ_2 coordinate, whereas if we move along the plane, we can change both the θ_1 and θ_1' coordinates in any proportion. The stratum S_{10} is the θ_2-axis with the origin removed; it is a set of points from which only one degree of freedom of motion along the fiber is available. S_{01} is the θ_1-θ_1' plane with the origin removed; from these points, two degrees of freedom are available. The subscripts on each S are $\operatorname{rk} W_2$ and $\operatorname{rk} W_1$.

One goal of this paper is to provide a basis for all possible degrees of freedom from a point $\theta \in \mu^{-1}(W)$—that is, a basis that can express the initial directions of all possible smooth paths on the fiber leaving θ. The tricky part comes if θ is a point where many strata meet. We ask that for every stratum S adjoining θ, exactly $\dim S$ of the vectors in θ's basis should suffice to span all possible directions by which a smooth path on S can leave θ. In Figure 2, the coordinate axes can serve that role at S_{00}. But in general, the vectors in the basis cannot always be orthogonal to each other, because strata do not always meet each other at right angles, and every point on $\mu^{-1}(W)$ may need a different basis, because fibers can be curvy.

The purple arcs in Figure 2 and Figure 3 depict *moves* along basis directions. For any point on the pink stratum S_{01} in Figure 2, the basis consists of one vector that does not change the rowspace of W_1 (hence it

[1] The strata are, in the language of topology, *manifolds without boundary*, as for every stratum S and every point $p \in S$, there is an open neighborhood $N \subset S$ that contains p and is homeomorphic to a ball of the same dimension as the stratum. Unfortunately, the term "boundary" has conflicting meanings in topology: the term "manifold without boundary" is defined in a fashion that takes S to be the entire topological space, with no larger context. However, when we consider S as a point set in the topological space $\mathbb{R}^{d_\theta}$, the *boundary* of S is defined to be the set of points that lie in both the closure of S and the closure of $\mathbb{R}^{d_\theta} \setminus S$. But in our context, the dimension of a stratum S is always less than d_θ, so the closure of $\mathbb{R}^{d_\theta} \setminus S$ is $\mathbb{R}^{d_\theta}$. Therefore, the boundary of a stratum S is the closure of S, which is typically a strict superset of S. For example, in Figure 2, S_{01} is a plane with the origin removed, the closure of S_{01} is the whole plane—hence S_{00} lies in the closure of S_{01}—and the boundary of S_{01} also is the whole plane. So our manifolds without (intrinsic) boundary have (extrinsic) boundaries.

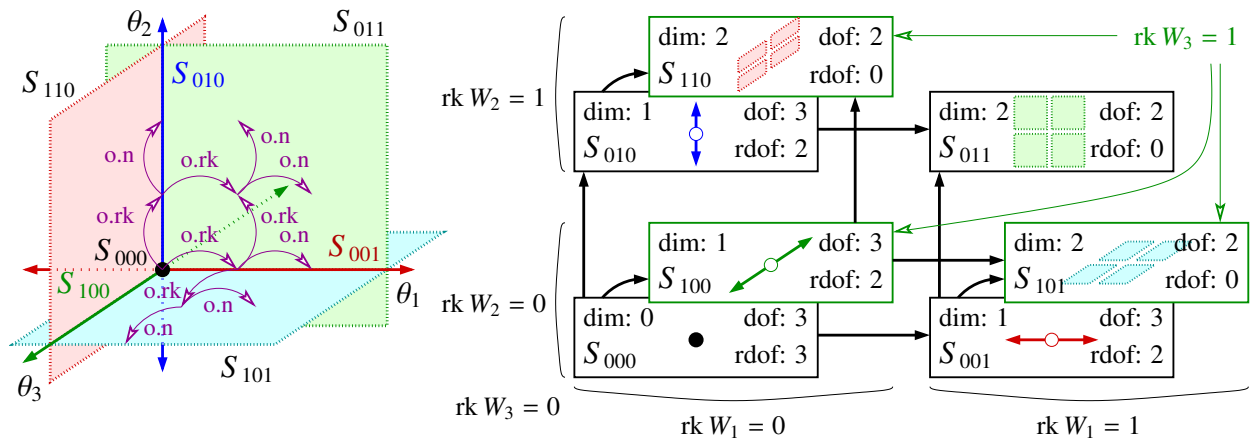


Figure 3: At left is the variety of solutions to $W_3W_2W_1 = [\theta_3][\theta_2][\theta_1] = [0] = W$, partitioned into seven strata: S_{000} is the origin; S_{001}, S_{010}, and S_{100} are the three coordinate axes with the origin removed; and S_{011}, S_{101}, and S_{110} are the three coordinate planes with the coordinate axes removed. The purple arcs show examples of two types of moves: a type o.rk move modifies one of the three matrices (W_3, W_2, or W_1) and increases its rank, thereby moving to a different stratum; whereas a type o.n move modifies one matrix without changing its rank nor leaving the current stratum. At right, the strata are arranged in a dag, which is organized as a three-dimensional table indexed by the ranks of W_3, W_2, and W_1.

points directly away from or toward the origin) and a second vector that does change the rowspace; together they span S_{01}. Three types of moves appear in the figure: *type o.rk* moves increase the rank of a matrix by one, thereby moving from one stratum to a higher-dimensional stratum; *type o.r* moves change a matrix's rowspace but not its columnspace, so the rank does not increase; and *type o.n* moves change neither a rowspace nor a columnspace. The prefix "o" denotes a *one-matrix move*, which changes only one factor matrix. There are two other types of moves, not relevant in this example: *type o.c* moves change some matrix's columnspace but not its rowspace, and *type t* moves are *two-matrix moves* that change two factor matrices simultaneously to negotiate the curvature of a stratum. (In Figure 1, for instance, all moves are of type t.) All of these moves stay on one stratum except the type o.rk moves, which provide us intuition for how the strata are connected to each other.

At right in Figure 2, we arrange the strata in a directed acyclic graph (dag) with the property that if the dag contains an edge (S_a, S_b), then the stratum S_a is a subset of the closure of S_b.[2] (The closure is taken with respect to the weight space $\mathbb{R}^{d_\theta}$.) For each stratum, the table lists the dimension of the stratum (dim), the number of degrees of freedom along which smooth motion on the fiber is possible (dof), and how many of those degrees of freedom increase a rank in the rank list (rdof, for "rank-increasing degrees of freedom").

Figure 3 depicts a second example, in which the fiber $\mu^{-1}([0])$ is the variety of solutions to $W_3W_2W_1 = [\theta_3][\theta_2][\theta_1] = [0] = W$. The dag showing how the seven strata are connected is a three-dimensional table: the strata are indexed by rk W_3, rk W_2, and rk W_1. Ordinarily, three-matrix fibers ($L = 3$) require five indices to index the strata, as rk W_3W_2 and rk W_2W_1 can vary as well; but as every matrix here is 1×1, those two ranks are uniquely determined by the first three. In the general case, the table is indexed by the rank list. If we leave out the ranks that don't change (the d_j's and rk W), the table has $L(L+1)/2 - 1$ dimensions and can have a very complicated shape.

Our third example is a fiber whose dimension is as high as 35 at some points, embedded in a 54-dimensional weight space. Table 1 depicts a dag that represents a stratification of the fiber $\mu^{-1}(W)$ for any 5×4 matrix

[2] If we replace the strata with their closures, then S_a is a subset of S_b and the dag is a Hasse diagram ordered by inclusion.

rk $W_2 = 5$	dim: 31 dof: 34 S_{51} rdof: 3	dim: 34 dof: 34 S_{52} rdof: 0		
rk $W_2 = 4$	dim: 29 dof: 37 S_{41} rdof: 8	dim: 33 dof: 36 S_{42} rdof: 3	dim: 35 dof: 35 S_{43} rdof: 0	
rk $W_2 = 3$	dim: 25 dof: 40 S_{31} rdof: 15	dim: 30 dof: 38 S_{32} rdof: 8	dim: 33 dof: 36 S_{33} rdof: 3	dim: 34 dof: 34 S_{34} rdof: 0
rk $W_2 = 2$	dim: 19 dof: 43 S_{21} rdof: 24	dim: 25 dof: 40 S_{22} rdof: 15	dim: 29 dof: 37 S_{23} rdof: 8	dim: 31 dof: 34 S_{24} rdof: 3
rk $W_2 = 1$	dim: 11 dof: 46 S_{11} rdof: 35	dim: 18 dof: 42 S_{12} rdof: 24	dim: 23 dof: 38 S_{13} rdof: 15	dim: 26 dof: 34 S_{14} rdof: 8
	rk $W_1 = 1$	rk $W_1 = 2$	rk $W_1 = 3$	rk $W_1 = 4$

Table 1: Dag representing the stratification of $\mu^{-1}(W)$ for $W_2 \in \mathbb{R}^{5\times 6}$, $W_1 \in \mathbb{R}^{6\times 4}$, and rk $W = 1$. The dag edges are omitted, but each stratum S_{ki} has an edge pointing to the stratum $S_{k+1,i}$ immediately above it, and another edge pointing to the stratum $S_{k,i+1}$ immediately to its right. The two points in weight space depicted in Figure 4 lie on the strata S_{32} and S_{33} in this table.

W with rank 1 and a network with $L = 2$, $d_2 = 5$, $d_1 = 6$, and $d_0 = 4$. (Two points on this fiber are illustrated in matrix form in Figure 4.)

In the two-matrix case ($L = 2$), the general shape for tables like Table 1 is a pentagon. Observe that the horizontal axis is rk W_1 and the vertical axis is rk W_2. The left and right boundaries of the table are determined by rk $W_1 \in [\text{rk } W, \min\{d_1, d_0\}]$, and the top and bottom boundaries are determined by rk $W_2 \in [\text{rk } W, \min\{d_2, d_1\}]$. Sometimes the upper right corner of the table is cut off by a fifth constraint: according to Sylvester's inequality, rk W_2 + rk $W_1 \leq d_1$ + rk W. This inequality generates the fifth edge of the pentagon.

One goal of this paper is to automate the generation of all the information in dags like Table 1 (given the d_j's and rk W), plus additional information such as how many rank-increasing degrees of freedom are associated with each dag edge, and a basis that describes all directions of motion on the fiber at a specified point.

Two-factor fibers ($L = 2$) are substantially easier to characterize than the general case; we summarize the moves and their degrees of freedom in Table 2 (without justification until Section 6). Starting from a point $\theta = (W_2, W_1) \in \mu^{-1}(W)$, a *move* on the fiber proceeds in the direction of a *displacement* $\Delta\theta = (\Delta W_2, \Delta W_1)$, chosen from one of the subspaces listed in the table. Each type of move has a different subspace to choose from. These subspaces are linearly independent of each other, so we can sum the degrees of freedom of each type to obtain the total number of degrees of freedom of motion from θ on the fiber ("dof"). At the bottom of Table 2, we give formulae for the degrees of freedom summarized as "dof", "rdof" (rank-increasing degrees of freedom), and "dim" (dimension of the stratum that contains θ) in Figure 2 and Table 1.

There are several caveats in interpreting Table 2, which are best appreciated by imagining that the moves are infinitesimal. First, we assume each move is sufficiently short that neither rk W_1 nor rk W_2 decreases. (An infinitesimal perturbation can increase the rank of a matrix, but not decrease it.) Second, the type t moves follow directions along which the fiber curves, so the displacement $(\Delta W_2, \Delta W_1)$ should be understood to be the *initial* direction of motion on a smooth, curved path on the fiber, not the displacement to the final destination. In Section 6, we explain the distinctions between *infinitesimal moves*, which are conceptually useful for understanding the dimensions of strata and how they are connected to each other, and *finite moves*, which are actual computations that change a neural network's weights.

Figure 4 illustrates how the nine different types of moves from Table 2 change the matrices W_2 and W_1, in the special case where the subspaces col W_2, row W_2, col W_1, and row W_1 (but not col W nor row W) are all

Move type	Displacement	Degrees of freedom
o.1	$\Delta W_1 \in \text{null}\, W_2 \otimes \mathbb{R}^{d_0}$; $\Delta W_2 = 0$	$(\omega_{10} + \omega_{11})\, d_0$, including:
o.n.1	$\Delta W_1 \in (\text{null}\, W_2 \cap \text{col}\, W_1) \otimes \text{row}\, W_1$	$\omega_{10}\, \text{rk}\, W_1$
o.c.1	$\Delta W_1 \in (\text{null}\, W_2 \cap (\text{null}\, W_2 \cap \text{col}\, W_1)^\perp) \otimes \text{row}\, W_1$	$\omega_{11}\, \text{rk}\, W_1$
o.r.1	$\Delta W_1 \in (\text{null}\, W_2 \cap \text{col}\, W_1) \otimes \text{null}\, W_1$	$\omega_{10}\, (d_0 - \text{rk}\, W_1)$
o.rk.1	$\Delta W_1 \in (\text{null}\, W_2 \cap (\text{null}\, W_2 \cap \text{col}\, W_1)^\perp) \otimes \text{null}\, W_1$	$\omega_{11}\, (d_0 - \text{rk}\, W_1)$
o.2	$\Delta W_2 \in \mathbb{R}^{d_2} \otimes \text{null}\, W_1^\top$; $\Delta W_1 = 0$	$(\omega_{21} + \omega_{11})\, d_2$, including:
o.n.2	$\Delta W_2 \in \text{col}\, W_2 \otimes (\text{row}\, W_2 \cap \text{null}\, W_1^\top)$	$\omega_{21}\, \text{rk}\, W_2$
o.r.2	$\Delta W_2 \in \text{col}\, W_2 \otimes ((\text{row}\, W_2 \cap \text{null}\, W_1^\top)^\perp \cap \text{null}\, W_1^\top)$	$\omega_{11}\, \text{rk}\, W_2$
o.c.2	$\Delta W_2 \in \text{null}\, W_2^\top \otimes (\text{row}\, W_2 \cap \text{null}\, W_1^\top)$	$\omega_{21}\, (d_2 - \text{rk}\, W_2)$
o.rk.2	$\Delta W_2 \in \text{null}\, W_2^\top \otimes ((\text{row}\, W_2 \cap \text{null}\, W_1^\top)^\perp \cap \text{null}\, W_1^\top)$	$\omega_{11}\, (d_2 - \text{rk}\, W_2)$
t	$\Delta W_2 = W_2 K$; $\Delta W_1 = -K W_1$; where $K \in \text{row}\, W_2 \otimes \text{col}\, W_1$ Note: $\Delta W_2 \in \text{col}\, W_2 \otimes \text{col}\, W_1$; $\Delta W_1 \in \text{row}\, W_2 \otimes \text{row}\, W_1$	$\text{rk}\, W_2 \cdot \text{rk}\, W_1$

Key: $\omega_{10} = \text{rk}\, W_1 - \text{rk}\, W$; $\omega_{21} = \text{rk}\, W_2 - \text{rk}\, W$; $\omega_{11} = d_1 - \text{rk}\, W_2 - \text{rk}\, W_1 + \text{rk}\, W$; $W = W_2 W_1$
Total degrees of freedom at θ: $\text{dof} = d_2 d_1 + d_1 d_0 - d_2\, \text{rk}\, W_1 - d_0\, \text{rk}\, W_2 + \text{rk}\, W_2 \cdot \text{rk}\, W_1$
Total rank-increasing degrees of freedom at θ: $\text{rdof} = \omega_{11}\, (d_2 + d_0 - \text{rk}\, W_2 - \text{rk}\, W_1)$
Dimension of θ's stratum: $\dim = \text{dof} - \text{rdof} = d_2 \omega_{21} + d_0 \omega_{10} + \omega_{11}\, (\text{rk}\, W_2 + \text{rk}\, W_1) + \text{rk}\, W_2 \cdot \text{rk}\, W_1$

Table 2: Summary of the types of moves and their degrees of freedom for a two-factor network ($L = 2$). Each move starts from a point $\theta = (W_2, W_1) \in \mu^{-1}(W)$ and moves in the direction of a displacement $\Delta\theta = (\Delta W_2, \Delta W_1)$. The tensor product $Y \otimes Z$ is the subspace of matrices $\{M : \text{col}\, M \subseteq Y \text{ and row}\, M \subseteq Z\}$.

axis-aligned. This axis alignment permits us to "see" the subspaces listed in Table 2 and count the degrees of freedom of motion along the fiber of W from (W_2, W_1). In particular, every number in every green box in W_2 can be changed to arbitrary values without changing the product W (so long as W_1 does not change). The same statement holds for the green boxes in W_1 (so long as W_2 does not change). The red boxes in W_2 and W_1 are coupled: in the top half of Figure 4, the red boxes represent six degrees of freedom (not twelve). The bottom half of Figure 4 shows how the subspaces and degrees of freedom change after W_1 is modified by a rank-increasing move (type o.rk.1), moving us from a stratum S_{32} to a different stratum S_{33}.

4 Subspace Flow through a Linear Neural Network

One of the fundamental concepts of linear algebra is that of the nullspace of a matrix W: the set of vectors x such that $Wx = 0$. Given a linear neural network, we can refine the concept by asking, at what layer does a particular input vector x first *disappear*? Formally, what is the smallest i such that $W_i W_{i-1} \cdots W_2 W_1 x = \mathbf{0}$? This question is answered by inspecting the nullspaces of the "part way there" matrices $W_{i\sim 0} = W_i W_{i-1} \cdots W_2 W_1$, which form a hierarchy of subspaces

$$\text{null}\, W \supseteq \text{null}\, W_{L-1\sim 0} \supseteq \text{null}\, W_{L-2\sim 0} \supseteq \ldots \supseteq \text{null}\, W_3 W_2 W_1 \supseteq \text{null}\, W_2 W_1 \supseteq \text{null}\, W_1.$$

We can extend the concept further by observing that linear neural networks can have unused subspaces in inner layers—subspaces through which information could flow if it were present, but the earlier layers are not putting any information into those subspaces. If a left nullspace $\text{null}\, W_i^\top$ is not the trivial subspace $\{\mathbf{0}\}$ (for example, if unit layer i has more units than the previous layer $i - 1$), then the space $\mathbb{R}^{d_i}$ encoded by unit layer i has one or more "wasted" dimensions that carry no information about the input x. We ask: if

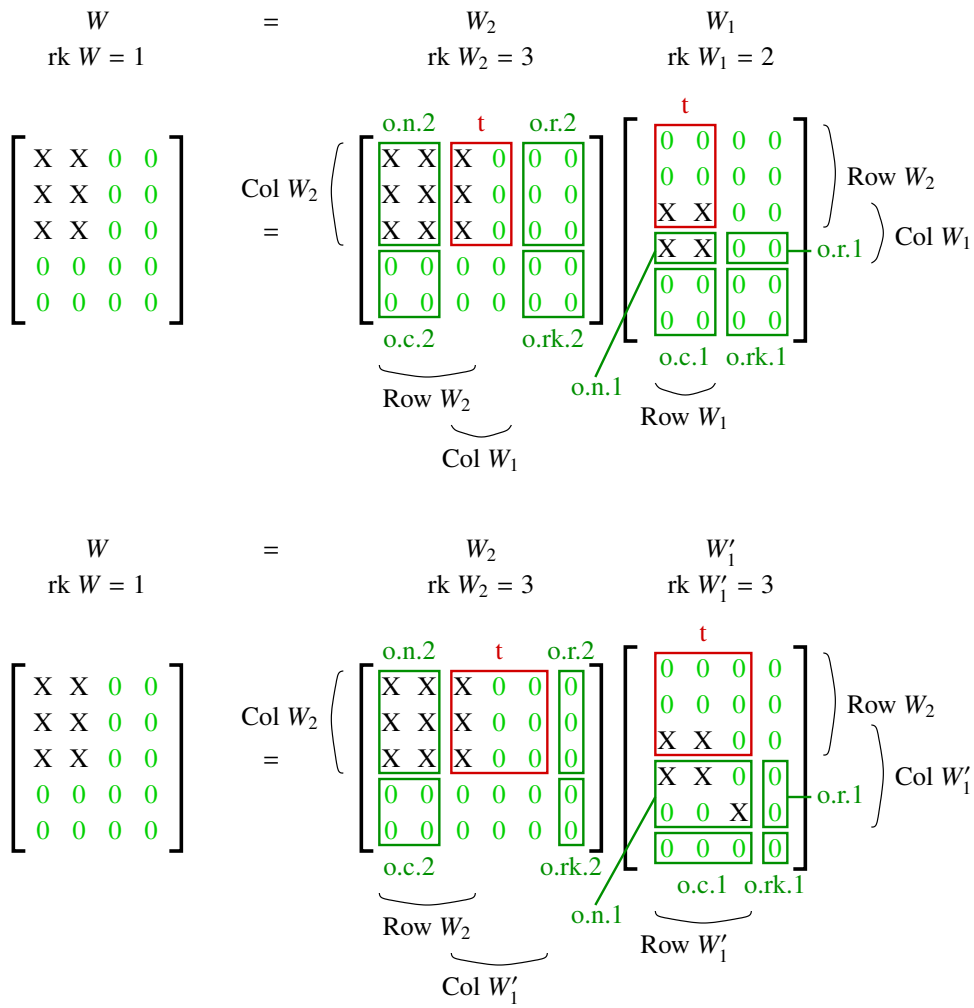


Figure 4: Top: a rank-1 matrix W as a product of a rank-3 matrix W_2 and a rank-2 matrix W_1. This factorization lies on the stratum S_{32} from Table 1. The X's represent arbitrary real values, so long as the matrices have the ranks specified. The red and green rectangles show which values can be changed by each of the nine different types of moves (o.n.2, o.c.2, o.r.2, o.rk.2, o.n.1, o.c.1, o.r.1, o.rk.1, or t). Bottom: another factorization of W found by applying a type o.rk.1 move to the top network, thereby replacing W_1 with a rank-3 matrix W_1' such that $W = W_2 W_1'$. This factorization lies on the stratum S_{33} from Table 1.

information were somehow injected into these left nullspaces, would it affect the network's output, or would it be absorbed in the nullspaces of subsequent matrices downstream? The answer is relevant to gradient descent algorithms for learning.

These questions are motivated by both practice and theory. The main practical motivation comes from neural network training. Although the wasted dimensions emerging from left nullspaces have no influence on the linear transformation W that the network computes, they have tremendous influence on whether a gradient descent algorithm can find weight updates that improve the network's performance. To illustrate this fact, consider the neural network with weight vector $\theta = \left(\begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix}, \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix}, \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix}\right)$. This network computes a linear transformation W of rank 1 and standard gradient descent algorithms cannot find a search direction that increases the rank above 1. Whereas in the network with weight vector

$\theta = \left(\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}, \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}, \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix} \right)$, which computes the same transformation, the subspace null $W_1^\top$ in hidden layer 1 is already connected to the network's output layer, so gradient descent can easily find a way to update W_1 that increases the ranks of both W_1 and W.

The main theoretical motivation arises because the fiber $\mu^{-1}(W)$ of a matrix W is not generally a manifold, but it can be written as a partition into strata, manifolds of various dimensions. We will define one stratum for each specific state of subspace flow through the network. These different "states of subspace flow" depend only on the rank list, so they are finite and combinatorial in nature, even though the transformations that the subspaces undergo are continuous and numerical in nature.

4.1 Interval Multisets and the Basis Flow Diagram

We will see that this state of information flow can be represented as a multiset of intervals, depicted in Figure 5. An *interval* is a set of consecutive integers $[i, k] = \{i, i+1, \ldots, k-1, k\}$ that identifies some consecutive unit layers in the network (with $0 \le i \le k \le L$). Each interval has a *multiplicity* ω_{ki}, representing ω_{ki} copies of the interval. If an interval is absent from the multiset, we say its multiplicity is zero (setting $\omega_{ki} = 0$).

Think of an interval $[i, k]$ with multiplicity ω_{ki} as representing an ω_{ki}-dimensional subspace that appears at unit layer i, being linearly independent of the columnspace of W_i (though not necessarily orthogonal to col W_i); then the subspace is linearly transformed by propagating through weight layers $W_{i+1}, W_{i+2}, \ldots, W_k$ to reach unit layer k with ω_{ki} dimensions still intact, only to disappear into the nullspace of W_{k+1} (unless layer k is the output layer). There is a second interpretation in terms of the transpose network $W^\top = W_1^\top W_2^\top \cdots W_{L-1}^\top W_L^\top$: the interval $[i, k]$ represents a (different!) ω_{ki}-dimensional subspace that appears at unit layer k, being linearly independent of the rowspace of W_{k+1}; then it is transformed by propagating through weight layers $W_k^\top, W_{k-1}^\top, \ldots, W_{i+1}^\top$ to reach unit layer i with ω_{ki} dimensions still intact, only to disappear into the left nullspace of W_i (if $i \neq 0$). We will need both interpretations to understand moves on the fiber $\mu^{-1}(W)$ and moves on each stratum.

A multiset of intervals is fully specified by the parameters $\omega_{ki} \ge 0$ for all k and i satisfying $L \ge k \ge i \ge 0$. A multiset of intervals is *valid* for a specified network if it satisfies the constraint that for each unit layer $j \in [0, L]$, the sum of the multiplicities of the intervals that contain j is d_j; that is,

$$d_j = \sum_{m=j}^{L} \sum_{l=0}^{j} \omega_{ml}. \tag{1}$$

Refer to Figure 5: you can see that verifying whether a multiset of intervals is valid is a simple matter of counting multiplicities in each unit layer. (We will see that the multiplicity ω_{ki} symbolizes ω_{ki} basis vectors for each unit layer $j \in [i, k]$, and the full set of d_j basis vectors at layer j is a basis for $\mathbb{R}^{d_j}$.) Therefore, only finitely many multisets are possible for a network with fixed layer sizes d_j. Each weight vector $\theta \in \mathbb{R}^{d_\theta}$ is associated with one multiset of intervals that expresses the subspace flows induced by θ, and θ lies on one stratum associated with that multiset.

Recall the *subsequence matrices* $W_{k\sim i} = W_k W_{k-1} \cdots W_{i+1}$. We will see in Section 4.5 that a multiset of intervals gives us an easy way to determine the rank of any subsequence matrix: the rank of $W_{k\sim i}$ is the total multiplicity of the intervals that contain both i and k. That is,

$$\operatorname{rk} W_{k\sim i} = \sum_{m=k}^{L} \sum_{l=0}^{i} \omega_{ml}. \tag{2}$$

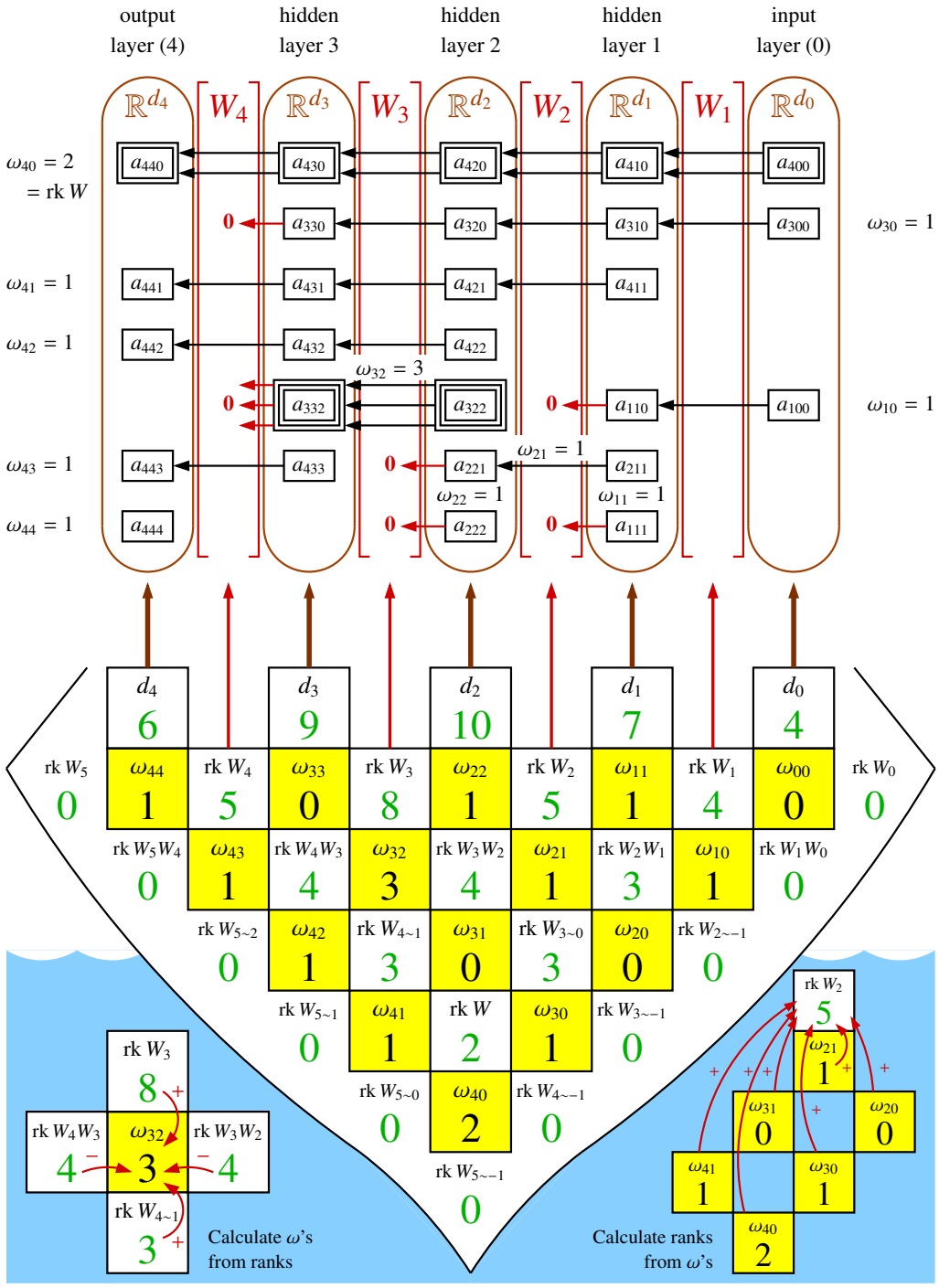


Figure 5: The tea clipper ship Basis Flow. The top half is a basis flow diagram that illustrates the flow of the prebasis subspaces a_{kji} through the network. Double boxes represent subspaces of dimension 2 and triple boxes represent subspaces of dimension 3. The bottom half shows the relationships between the intervals, the layer sizes, and the matrix ranks. The number of units d_j in unit layer j equals the sum of the multiplicities ω_{ml} of the intervals that touch layer j (i.e., the dimensions of the prebasis subspaces a_{mjl}). Each matrix rank $\operatorname{rk} W_{k\sim i}$ is the sum of the multiplicities of the intervals that touch both layers k and i.

Refer again to Figure 5: you can easily read the rank of each subsequence matrix off the intervals.

In particular,

$$\operatorname{rk} W = \omega_{L0}. \tag{3}$$

That is, the interval $[0, L]$ always has multiplicity $\operatorname{rk} W$; this interval represents the fact that the subspace $\operatorname{row} W$ at the input layer is mapped by W to $\operatorname{col} W$ at the output layer, and both subspaces have dimension $\operatorname{rk} W$. When we examine the fiber of a specific matrix W, we fix the rank $\operatorname{rk} W$ and the multiplicity ω_{L0}, as well as $W_{j\sim j} = d_j$ for $j \in [0, L]$. The other ranks and multiplicities generally vary across different factorizations of W.

Recall the *rank list* defined in Section 2. We will see in Section 4.5 that there is a bijection between valid multisets of intervals and valid rank lists: if we are given a list of interval multiplicities, we can easily determine the ranks, and if we are given a rank list, we can easily determine the interval multiplicities. A rank list is *valid* if there is some weight vector $\theta \in \mathbb{R}^{d_\theta}$ that achieves the matrix ranks listed. (We will see that the invalid rank lists are those with a negative rank and those that imply that one of the multiplicities is negative.) Our stratification of the fiber $\mu^{-1}(W)$ has one stratum for each valid multiset such that $\omega_{L0} = \operatorname{rk} W$; equivalently, one stratum for each valid rank list having the specified values of $\operatorname{rk} W$ and the d_j's.

4.2 Flow Subspaces and Subspace Hierarchies

In this section, we identify subspaces in each unit layer's space $\mathbb{R}^{d_j}$ that represent information flowing through the linear neural network (or through the transpose network $W^\top = W_1^\top W_2^\top \cdots W_{L-1}^\top W_L^\top$), with special attention to information that does not reach the output layer. We are aided in this effort by the fact that, at a unit layer j in the network, the fundamental subspaces associated with the subsequence matrices are nested in hierarchies as follows.

$$\begin{aligned}
&\mathbb{R}^{d_j} = \operatorname{row} W_{j\sim j} \supseteq \operatorname{row} W_{j+1} \supseteq \operatorname{row} W_{j+2}W_{j+1} \supseteq \ldots \supseteq \operatorname{row} W_{L\sim j} \supseteq \operatorname{row} W_{L+1\sim j} = \{\mathbf{0}\},\\
&\{\mathbf{0}\} = \operatorname{null} W_{j\sim j} \subseteq \operatorname{null} W_{j+1} \subseteq \operatorname{null} W_{j+2}W_{j+1} \subseteq \ldots \subseteq \operatorname{null} W_{L\sim j} \subseteq \operatorname{null} W_{L+1\sim j} = \mathbb{R}^{d_j},\\
&\mathbb{R}^{d_j} = \operatorname{col} W_{j\sim j} \supseteq \operatorname{col} W_j \supseteq \operatorname{col} W_jW_{j-1} \supseteq \ldots \supseteq \operatorname{col} W_{j\sim 0} \supseteq \operatorname{col} W_{j\sim -1} = \{\mathbf{0}\}, \quad \text{and}\\
&\{\mathbf{0}\} = \operatorname{null} W_{j\sim j}^\top \subseteq \operatorname{null} W_j^\top \subseteq \operatorname{null} (W_jW_{j-1})^\top \subseteq \ldots \subseteq \operatorname{null} W_{j\sim 0}^\top \subseteq \operatorname{null} W_{j\sim -1}^\top = \mathbb{R}^{d_j}.
\end{aligned}$$

By the Fundamental Theorem, the subspaces in the first row are orthogonal complements of the corresponding subspaces in the second row, and the subspaces in the third row are orthogonal complements of the corresponding subspaces in the fourth row. Here, we are using the following conventions for subsequence matrices.

$$\begin{aligned}
&W_{j\sim j} = I_{d_j\times d_j} \quad \text{(the } d_j \times d_j \text{ identity matrix).}\\
&\text{Hence, } \operatorname{row} W_{j\sim j} = \mathbb{R}^{d_j} = \operatorname{col} W_{j\sim j} \text{ and } \operatorname{null} W_{j\sim j} = \{\mathbf{0}\} = \operatorname{null} W_{j\sim j}^\top.\\
&W_{j\sim -1} = 0_{d_j\times 1} \quad \text{and} \quad W_{L+1\sim j} = 0_{1\times d_j} \quad \text{(zero matrices).}\\
&\text{Hence, } \operatorname{row} W_{L+1\sim j} = \{\mathbf{0}\} = \operatorname{col} W_{j\sim -1} \quad \text{and} \quad \operatorname{null} W_{L+1\sim j} = \mathbb{R}^{d_j} = \operatorname{null} W_{j\sim -1}^\top.
\end{aligned}$$

(Note that the last two lines are consistent with imagining that the network $W_LW_{L-1}\cdots W_1$ is sandwiched between two extra matrices $W_{L+1} = 0$ and $W_0 = 0$.)

From these four hierarchies, we define two hierarchies of *flow subspaces* that give us insight about how information flows, and sometimes fails to flow, through the network. The flow subspaces of $\mathbb{R}^{d_j}$ at unit layer

$j \in [0, L]$ are

$$A_{kji} = \text{null}\, W_{k+1\sim j} \cap \text{col}\, W_{j\sim i}, \qquad i \in [-1, j], k \in [j-1, L], \qquad \text{and}$$
$$B_{kji} = \text{row}\, W_{k\sim j} \cap \text{null}\, W^\top_{j\sim i-1}, \qquad i \in [0, j+1], k \in [j, L+1].$$

For example, $A_{320} = \text{null}\, W_4 W_3 \cap \text{col}\, W_2 W_1$ and $B_{320} = \text{row}\, W_3 \cap \text{null}\, W^\top_{2\sim -1} = \text{row}\, W_3$. We will use commas to separate the subscripts when necessary for clarity; e.g., $A_{y-1,x+1,-1}$. Intuitively, $A_{kji} \in \mathbb{R}^{d_j}$ is the subspace that carries information in unit layer j that has come at least as far as from layer i, but will not survive farther than layer k. In the transpose network $W^\top = W_1^\top W_2^\top \cdots W_{L-1}^\top W_L^\top$, $B_{kji} \in \mathbb{R}^{d_j}$ is the subspace that carries information in unit layer j that has come at least as far as from layer k, but will not survive farther than layer i.

We need a notation for the dimensions of the flow subspaces. Let

$$\alpha_{kji} = \dim A_{kji} \quad \text{and} \quad \beta_{kji} = \dim B_{kji}.$$

It is easy to see that

$$A_{kji} \supseteq A_{k'ji'} \text{ and } \alpha_{kji} \geq \alpha_{k'ji'} \quad \text{if} \quad k \geq k' \text{ and } i \geq i', \text{ assuming } j \in [0, L], k, k' \in [j-1, L], i, i' \in [-1, j].$$
$$B_{kji} \subseteq B_{k'ji'} \text{ and } \beta_{kji} \leq \beta_{k'ji'} \quad \text{if} \quad k \geq k' \text{ and } i \geq i', \text{ assuming } j \in [0, L], k, k' \in [j, L+1], i, i' \in [0, j+1].$$

Table 3 depicts this relationship and the partial ordering it imposes on the flow subspaces.

Let us consider the relationships between flow subspaces at different unit layers of the network. Given a matrix W and a subspace A, we define

$$WA = \{Wv : v \in A\},$$

which is also a subspace. The simplest flow relationships are that $A_{kji} = W_j A_{k,j-1,i}$ and $B_{k,j-1,i} = W_j^\top B_{kji}$, which exposes why we call them *flow subspaces*: you may imagine the A subspaces flowing through the network, being linearly transformed layer by layer; and you may imagine the B subspaces flowing through the transpose network $W^\top = W_1^\top W_2^\top \cdots W_{L-1}^\top W_L^\top$, also being transformed at each layer. Figure 6 depicts flow subspaces at each unit layer of a linear neural network. The following lemma expresses these relationships in a slightly more general way.

Lemma 1. *$A_{kji} = W_{j\sim x} A_{kxi}$ for all k, j, i, and x that satisfy $L \geq k$ and $k+1 \geq j \geq x \geq i \geq 0$.*

Furthermore, $B_{kji} = W^\top_{y\sim j} B_{kyi}$ for all k, j, i, and y that satisfy $L \geq k \geq y \geq j \geq i-1$ and $i \geq 0$.

Proof. By definition, $A_{kii} = \text{null}\, W_{k+1\sim i} \cap \mathbb{R}^{d_i} = \text{null}\, W_{k+1\sim i}$. Hence $W_{k+1\sim i} A_{kii} = \{\mathbf{0}\}$. For every $z \in [i, k+1]$, $W_{k+1\sim z} W_{z\sim i} A_{kii} = \{\mathbf{0}\}$ and thus $W_{z\sim i} A_{kii} \subseteq \text{null}\, W_{k+1\sim z}$. Obviously, $W_{z\sim i} A_{kii} \subseteq \text{col}\, W_{z\sim i}$. Hence $W_{z\sim i} A_{kii} \subseteq \text{null}\, W_{k+1\sim z} \cap \text{col}\, W_{z\sim i} = A_{kzi}$.

To see that the reverse inclusion also holds, consider a vector $v \in A_{kzi}$. As $v \in \text{col}\, W_{z\sim i}$, there is a vector $w \in \mathbb{R}^{d_i}$ such that $v = W_{z\sim i} w$. As $v \in \text{null}\, W_{k+1\sim z}$, we have $\mathbf{0} = W_{k+1\sim z} v = W_{k+1\sim z} W_{z\sim i} w = W_{k+1\sim i} w$, so $w \in \text{null}\, W_{k+1\sim i} = A_{kii}$ and thus $v \in W_{z\sim i} A_{kii}$. Hence $W_{z\sim i} A_{kii} \supseteq A_{kzi}$; hence $W_{z\sim i} A_{kii} = A_{kzi}$ for every $z \in [i, k+1]$.

It follows that

$$\begin{aligned} W_{j\sim i} A_{kii} &= W_{j\sim x} W_{x\sim i} A_{kii} \quad \text{and} \\ A_{kji} &= W_{j\sim x} A_{kxi} \end{aligned}$$

as claimed. Applying the same proof to the transpose network shows that $B_{kji} = W^\top_{y\sim j} B_{kyi}$. □

$\mathbb{R}^{d_2}$	$k=4$	$A_{4,2,-1}=\{\mathbf{0}\}$	$\subseteq$	A_{420}	$\subseteq$	A_{421}	$\subseteq$	$A_{422}=\mathbb{R}^{d_2}$
$\cup\!\!\mid$				$\cup\!\!\mid$		$\cup\!\!\mid$		$\cup\!\!\mid$
null W_4W_3	$k=3$	$A_{3,2,-1}=\{\mathbf{0}\}$	$\subseteq$	A_{320}	$\subseteq$	A_{321}	$\subseteq$	A_{322}
$\cup\!\!\mid$				$\cup\!\!\mid$		$\cup\!\!\mid$		$\cup\!\!\mid$
null W_3	$k=2$	$A_{2,2,-1}=\{\mathbf{0}\}$	$\subseteq$	A_{220}	$\subseteq$	A_{221}	$\subseteq$	A_{222}
$\cup\!\!\mid$				$\cup\!\!\mid$		$\cup\!\!\mid$		$\cup\!\!\mid$
$\{\mathbf{0}\}$	$k=1$	$A_{1,2,-1}=\{\mathbf{0}\}$		$A_{120}=\{\mathbf{0}\}$		$A_{121}=\{\mathbf{0}\}$		$A_{122}=\{\mathbf{0}\}$
null $W_{k+1\sim 2}$		$i=-1$		$i=0$		$i=1$		$i=2$
$A_{k2i}\nearrow$	col $W_{2\sim i}$	$\{\mathbf{0}\}$	$\subseteq$	col W_2W_1	$\subseteq$	col W_2	$\subseteq$	$\mathbb{R}^{d_2}$

$\{\mathbf{0}\}$	$k=5$	$B_{520}=\{\mathbf{0}\}$		$B_{521}=\{\mathbf{0}\}$		$B_{522}=\{\mathbf{0}\}$		$B_{523}=\{\mathbf{0}\}$
$\cap\!\!\mid$		$\cap\!\!\mid$		$\cap\!\!\mid$		$\cap\!\!\mid$		
row W_4W_3	$k=4$	B_{420}	$\supseteq$	B_{421}	$\supseteq$	B_{422}	$\supseteq$	$B_{423}=\{\mathbf{0}\}$
$\cap\!\!\mid$		$\cap\!\!\mid$		$\cap\!\!\mid$		$\cap\!\!\mid$		
row W_3	$k=3$	B_{320}	$\supseteq$	B_{321}	$\supseteq$	B_{322}	$\supseteq$	$B_{323}=\{\mathbf{0}\}$
$\cap\!\!\mid$		$\cap\!\!\mid$		$\cap\!\!\mid$		$\cap\!\!\mid$		
$\mathbb{R}^{d_2}$	$k=2$	$B_{220}=\mathbb{R}^{d_2}$	$\supseteq$	B_{221}	$\supseteq$	B_{222}	$\supseteq$	$B_{223}=\{\mathbf{0}\}$
row $W_{k\sim 2}$		$i=0$		$i=1$		$i=2$		$i=3$
$B_{k2i}\nearrow$	null $W_{2\sim i-1}$	$\mathbb{R}^{d_2}$	$\supseteq$	null $(W_2W_1)^\top$	$\supseteq$	null $W_2^\top$	$\supseteq$	$\{\mathbf{0}\}$

Table 3: The hierarchical nesting of the flow subspaces at unit layer $j=2$ of a network with $L=4$ matrices. Top: $A_{k2i}=\text{null } W_{k+1\sim 2}\cap \text{col } W_{2\sim i}$ for each k,i. Bottom: $B_{k2i}=\text{row } W_{k\sim 2}\cap \text{null } W_{2\sim i-1}$ for each k,i.

4.3 Bases and "Prebases" for the Flow Subspaces

In this section, we show how to decompose each unit layer's space $\mathbb{R}^{d_j}$ into a "prebasis" of subspaces. We assume the reader is familiar with the standard idea from linear algebra of a *basis* for $\mathbb{R}^d$, comprising d linearly independent *basis vectors*. A prebasis is like a basis, but it is made up of subspaces rather than vectors; see below for a definition. Our prebasis for $\mathbb{R}^{d_j}$ includes (as a subset) a prebasis for every flow subspace A_{kji} (where the index j matches $\mathbb{R}^{d_j}$ but k and i vary freely). We also define a second prebasis for $\mathbb{R}^{d_j}$ that includes a prebasis for every flow subspace B_{kji} (with matching j); this prebasis represents subspaces that flow through the transpose network $W^\top = W_1^\top W_2^\top \cdots W_{L-1}^\top W_L^\top$.

Given two subspaces $X, Y \in \mathbb{R}^d$, their vector sum is $X+Y=\{x+y : x\in X \text{ and } y\in Y\}$. If X and Y are linearly independent—that is, if $X\cap Y=\{\mathbf{0}\}$—then $X+Y$ is called a *direct sum*, sometimes written $X\oplus Y$.[3] Likewise, given a set of subspaces $\mathcal{X}=\{X_1,X_2,\ldots,X_m\}$, the direct sum notation $X_1\oplus X_2\oplus\ldots\oplus X_m$ implies that the subspaces in $\mathcal{X}$ are linearly independent, meaning that for every $i\in[1,m]$, $X_i\cap\sum_{j\neq i}X_j=\{\mathbf{0}\}$.

If $\mathbb{R}^d = X_1\oplus X_2\oplus\ldots\oplus X_m$, then $\mathcal{X}=\{X_1,X_2,\ldots,X_m\}$ is known as a *direct sum decomposition* of $\mathbb{R}^d$. That's too many syllables, so we will call $\mathcal{X}$ a *prebasis* for $\mathbb{R}^d$ throughout this paper. We call each X_i a *prebasis subspace*. The linear independence of the prebasis subspaces implies that for every vector $v\in\mathbb{R}^d$, there is only one way to express v as a sum of vectors $v=\sum_{i=1}^m v_i$ such that $v_i\in X_i$. It also implies that $d=\dim X_1+\dim X_2+\ldots+\dim X_m$. If desired, it is conceptually easy to convert a prebasis into a traditional

[3]The notation $X\oplus Y$ is weird, because as an operator it produces exactly the same result as $X+Y$, but the operator notation itself implies a constraint on the subspaces X and Y: that $X\cap Y=\{\mathbf{0}\}$. If $X\cap Y\neq\{\mathbf{0}\}$, $X\oplus Y$ is undefined.

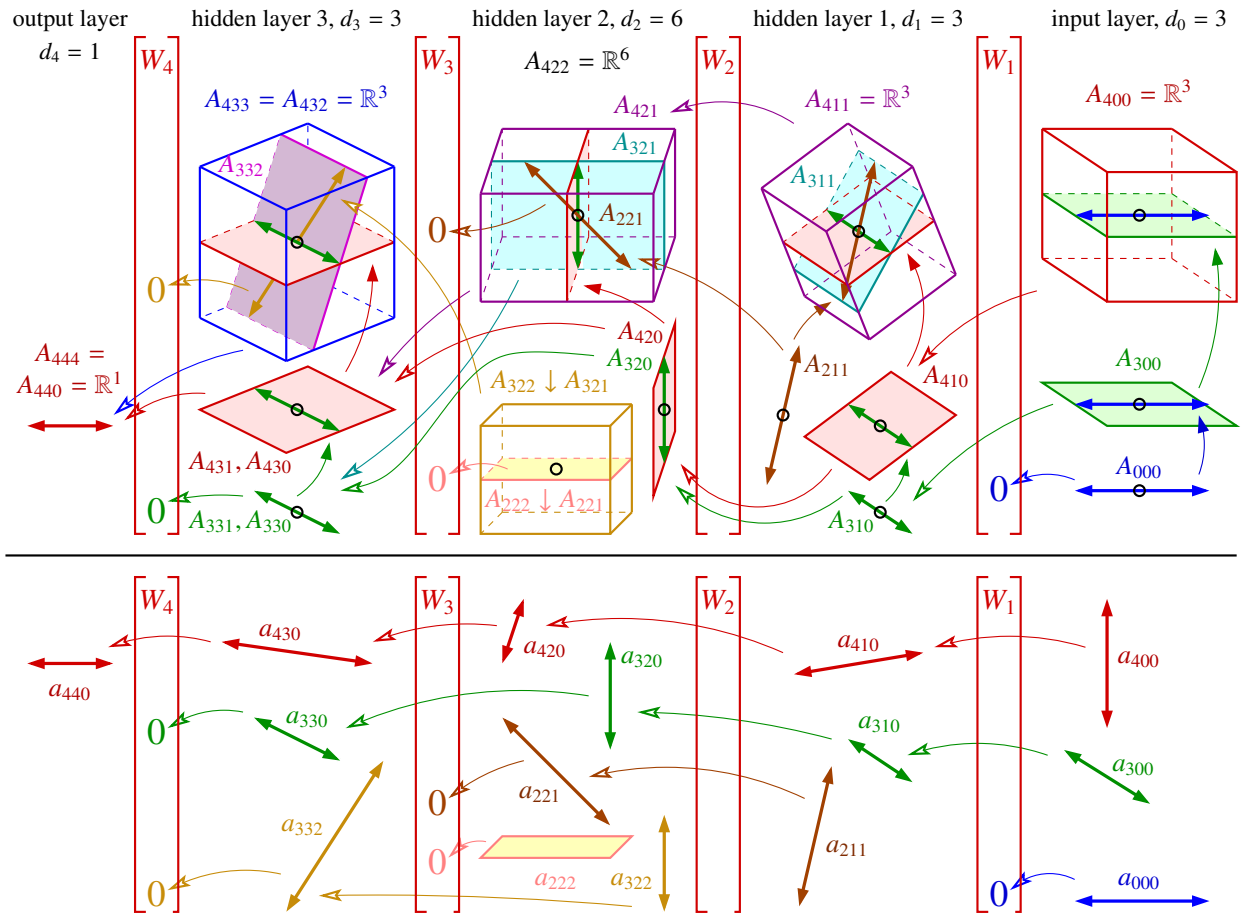


Figure 6: Top: an example of flow subspaces A_{kji}. Note that the subspace A_{322} is five-dimensional, so we cannot easily draw it complete. Instead, we draw $A_{322} \downarrow A_{321}$, which is a three-dimensional subspace of A_{322} linearly independent of A_{321}, and we draw the plane A_{321} separately; A_{322} is the vector sum of $A_{322} \downarrow A_{321}$ and A_{321}. Similarly, we draw $A_{222} \downarrow A_{221}$, a two-dimensional subspace of A_{222} linearly independent of A_{221}. A_{222} is a three-dimensional subspace, the vector sum of the plane $A_{222} \downarrow A_{221}$ and the line A_{221}. Bottom: an example of corresponding prebasis subspaces a_{kji}, forming a flow prebasis.

vector basis: just choose a basis for each X_i, then pool the d vectors together to form a basis for $\mathbb{R}^d$—hence the name "prebasis." Why don't we do that here? Because details like the choice of basis for each prebasis subspace and the length of each basis vector are irrelevant to our account and would make our presentation more complicated.

We define a custom operator to help us choose a prebasis. Given two vector subspaces $Y \subseteq Z$, we define the set of subspaces

$$Z \downarrow Y = \{X \subseteq Z : Z = X \oplus Y\}.$$

These subspaces all have the same dimension, namely, $\dim Z - \dim Y$, and they are all linearly independent of Y. There are two special cases where $Z \downarrow Y$ contains only one element: if $Y = \{\mathbf{0}\}$ then $Z \downarrow Y = \{Z\}$, and if $Y = Z$ then $Z \downarrow Y = \{\{\mathbf{0}\}\}$. Otherwise, $Z \downarrow Y$ is an infinite set of subspaces.

Recall the flow subspaces A_{kji} and B_{kji} from Section 4.2, both of them subspaces of $\mathbb{R}^{d_j}$, and recall that $A_{k,j,i-1} \subseteq A_{kji}$ and $A_{k-1,j,i} \subseteq A_{kji}$, assuming $L \geq k \geq j \geq i \geq 0$. It follows that $A_{k,j,i-1} + A_{k-1,j,i} \subseteq A_{kji}$.

Symmetrically, $B_{k,j,i+1} + B_{k+1,j,i} \subseteq B_{kji}$. For all i, j, and k satisfying $L \geq k \geq j \geq i \geq 0$, we choose *prebasis subspaces*

$$\begin{aligned} a_{kji} &\in A_{kji} \downarrow (A_{k,j,i-1} + A_{k-1,j,i}) \quad \text{and} \\ b_{kji} &\in B_{kji} \downarrow (B_{k,j,i+1} + B_{k+1,j,i}). \end{aligned}$$

It is common that some of these prebasis subspaces are simply $\{\mathbf{0}\}$; these can be omitted from any prebasis. When applying these definitions, recall that $A_{k,j,-1} = A_{j-1,j,i} = B_{k,j,j+1} = B_{L+1,j,i} = \{\mathbf{0}\}$ (so for example, $a_{jj0} = A_{jj0}$ and $b_{Ljj} = B_{Ljj}$). The bottom half of Figure 6 shows examples of prebasis subspaces chosen from these sets.

One element in $Z \downarrow Y$ is the subspace containing every vector in Z that is orthogonal to every vector in Y (written $Z \cap Y^{\perp}$), and it is tempting to always choose that subspace when we choose a_{kji} and b_{kji}, yielding what we call *standard prebases*. However, in Section 4.4 we exploit the flexibility that $Z \downarrow Y$ gives us to choose *flow prebases* instead, so the prebasis subspaces (a's and b's) "flow" through the network as the flow subspaces (A's and B's) do, as Figures 5 and 6 depict.

Lemma 3 below states that $\dim a_{kji} = \dim b_{kji}$, a crucial result that surprised us when we stumbled upon it. This establishes a pleasing symmetry between flow through a linear neural network and flow through its transpose network, even though the flow subspaces and their prebases are different. In Figure 5, we could depict the flow of prebasis subspaces through the transpose neural network simply by replacing each a_{kji} by b_{kji} and reversing the directions of the arrows in the top half of the figure. Lemma 3 also shows that the dimensions of the prebasis subspaces are easily computed from the dimensions of the flow subspaces. (The dimensions of the prebasis subspaces do not depend on which ones we choose.)

Two subspaces Y and Z are *orthogonal* if for every vector $y \in Y$ and every $z \in Z$, $y^{\top} z = 0$. The *orthogonal complement* of a subspace $Z \in \mathbb{R}^d$, denoted $Z^{\perp}$, is the set of vectors in $\mathbb{R}^d$ that are orthogonal to every vector in Z. Orthogonal complements have complementary dimensions: $\dim Z + \dim Z^{\perp} = d$. Linear algebra furnishes two classic examples: $(\operatorname{row} W)^{\perp} = \operatorname{null} W$ and $(\operatorname{col} W)^{\perp} = \operatorname{null} W^{\top}$. The following lemma prepares us for Lemma 3.

Lemma 2. *Consider subspaces $J \subseteq K \subseteq \mathbb{R}^d$ and $Y \subseteq Z \subseteq \mathbb{R}^d$. Then*

$$\begin{aligned} &\dim(K \cap Z) - \dim(K \cap Y + J \cap Z) \\ &= \dim(J^{\perp} \cap Y^{\perp}) - \dim(K^{\perp} \cap Y^{\perp} + J^{\perp} \cap Z^{\perp}) \\ &= \dim(K \cap Z) - \dim(K \cap Y) - \dim(J \cap Z) + \dim(J \cap Y) \\ &= \dim(J^{\perp} \cap Y^{\perp}) - \dim(K^{\perp} \cap Y^{\perp}) - \dim(J^{\perp} \cap Z^{\perp}) + \dim(K^{\perp} \cap Z^{\perp}). \end{aligned}$$

Proof. It is a property of vector subspaces that $\dim(E+F) + \dim(E \cap F) = \dim E + \dim F$. Letting $E = K \cap Y$ and $F = J \cap Z$, we have $E \cap F = J \cap Y$, which explains why the first expression equals the third one. Letting $E = K^{\perp} \cap Y^{\perp}$ and $F = J^{\perp} \cap Z^{\perp}$, we have $E \cap F = K^{\perp} \cap Z^{\perp}$, which explains why the second expression equals the fourth one.

To verify that the third expression equals the fourth one, we also use the De Morgan laws $(E+F)^{\perp} = E^{\perp} \cap F^{\perp}$

and $(E \cap F)^\perp = E^\perp + F^\perp$.

$$\begin{aligned}
&\dim(J^\perp \cap Y^\perp) - \dim(K^\perp \cap Y^\perp) - \dim(J^\perp \cap Z^\perp) + \dim(K^\perp \cap Z^\perp) \\
&= \dim J^\perp + \dim Y^\perp - \dim(J^\perp + Y^\perp) - \dim K^\perp - \dim Y^\perp + \dim(K^\perp + Y^\perp) \\
&\quad - \dim J^\perp - \dim Z^\perp + \dim(J^\perp + Z^\perp) + \dim(K^\perp \cap Z^\perp) \\
&= -\dim(J \cap Y)^\perp - \dim K^\perp + \dim(K \cap Y)^\perp - \dim Z^\perp + \dim(J \cap Z)^\perp + \dim(K + Z)^\perp \\
&= -d + \dim(J \cap Y) - d + \dim K + d - \dim(K \cap Y) - d + \dim Z + d - \dim(J \cap Z) + d - \dim(K + Z) \\
&= \dim(K \cap Z) - \dim(K \cap Y) - \dim(J \cap Z) + \dim(J \cap Y).
\end{aligned}$$

□

Lemma 3. *For $L \geq k \geq j \geq i \geq 0$, $\dim a_{kji} = \dim b_{kji} = \alpha_{kji} - \alpha_{k,j,i-1} - \alpha_{k-1,j,i} + \alpha_{k-1,j,i-1} = \beta_{kji} - \beta_{k,j,i+1} - \beta_{k+1,j,i} + \beta_{k+1,j,i+1}$ (recalling that $\alpha_{kji} = \dim A_{kji}$ and $\beta_{kji} = \dim B_{kji}$).*

Proof. As $a_{kji} \in A_{kji} \downarrow (A_{k,j,i-1} + A_{k-1,j,i})$, it follows from the definition of the operator $\downarrow$ that $\dim a_{kji} = \dim A_{kji} - \dim(A_{k,j,i-1} + A_{k-1,j,i})$. Similarly, $\dim b_{kji} = \dim B_{kji} - \dim(B_{k,j,i+1} + B_{k+1,j,i})$. The result follows from Lemma 2 by substituting $K = \operatorname{null} W_{k+1\sim j}$, $J = \operatorname{null} W_{k\sim j}$, $Z = \operatorname{col} W_{j\sim i}$, and $Y = \operatorname{col} W_{j\sim i-1}$. (Then $A_{kji} = K \cap Z$, $A_{k,j,i-1} = K \cap Y$, $A_{k-1,j,i} = J \cap Z$, $A_{k-1,j,i-1} = J \cap Y$, $B_{kji} = J^\perp \cap Y^\perp$, $B_{k,j,i+1} = J^\perp \cap Z^\perp$, $B_{k+1,j,i} = K^\perp \cap Y^\perp$, and $B_{k+1,j,i+1} = K^\perp \cap Z^\perp$.) □

We define two classes of prebases to span the subspaces A_{kji} and B_{kji}. Let $\mathcal{A}_{kji}$ be the set containing the subspaces $a_{k'ji'}$ for all $k' \in [j,k]$, $i' \in [0,i]$. Let $\mathcal{B}_{kji}$ be the set containing the subspaces $b_{k'ji'}$ for all $k' \in [k,L]$, $i' \in [i,j]$.

Lemma 4. *Given that $L \geq k \geq j \geq i \geq 0$, $\mathcal{A}_{kji}$ is a prebasis for A_{kji} and $\mathcal{B}_{kji}$ is a prebasis for B_{kji}.*

Proof. We prove the first claim by induction on increasing values of k and i. For the base cases, recall our convention that $A_{k,j,-1} = \{\mathbf{0}\}$ and $A_{j-1,j,i} = \{\mathbf{0}\}$. The empty set is a prebasis for the subspace $\{\mathbf{0}\}$, so we establish a convention that $\mathcal{A}_{k,j,-1} = \emptyset$ and $\mathcal{A}_{j-1,j,i} = \emptyset$.

For the inductive case—showing that $\mathcal{A}_{kji}$ is a prebasis for A_{kji}—we assume the inductive hypothesis that $\mathcal{A}_{k,j,i-1}$ is a prebasis for $A_{k,j,i-1}$, $\mathcal{A}_{k-1,j,i}$ is a prebasis for $A_{k-1,j,i}$, and $\mathcal{A}_{k-1,j,i-1}$ is a prebasis for $A_{k-1,j,i-1}$. Most of the work in this proof is to show that $\mathcal{A}_{k,j,i-1} \cup \mathcal{A}_{k-1,j,i}$ is a prebasis for $A_{k,j,i-1} + A_{k-1,j,i}$. Clearly, $A_{k,j,i-1} + A_{k-1,j,i}$ equals the vector sum of the subspaces in $\mathcal{A}_{k,j,i-1} \cup \mathcal{A}_{k-1,j,i}$. But we must also show that the subspaces in $\mathcal{A}_{k,j,i-1} \cup \mathcal{A}_{k-1,j,i}$ are linearly independent of each other.

Suppose for the sake of contradiction that they are linearly dependent. Then there exists a nonempty set V of nonzero vectors in $\mathbb{R}^{d_j}$ with sum zero such that each vector in V comes from a different subspace in $\mathcal{A}_{k,j,i-1} \cup \mathcal{A}_{k-1,j,i}$. Partition V into two disjoint subsets V' and V'' such that each vector in V' comes from a different subspace in $\mathcal{A}_{k,j,i-1}$ and each vector in V'' comes from a different subspace in $\mathcal{A}_{k-1,j,i} \setminus \mathcal{A}_{k,j,i-1}$. Let w be the sum of the vectors in V'. The sum of the vectors in V'' is $-w$. As V is nonempty, at least one of V' or V'' is nonempty. As the vectors in V' come from a prebasis ($\mathcal{A}_{k,j,i-1}$) and the vectors in V'' come from a prebasis ($\mathcal{A}_{k-1,j,i}$), $w \neq \mathbf{0}$ and both V' and V'' are nonempty. The vectors in V' are all in the subspace $A_{k,j,i-1}$, so $w \in A_{k,j,i-1}$; and the vectors in V'' are all in $A_{k-1,j,i}$, so $w \in A_{k-1,j,i}$. Therefore, $w \in A_{k,j,i-1} \cap A_{k-1,j,i} = \operatorname{null} W_{k\sim j} \cap \operatorname{col} W_{j\sim i-1} = A_{k-1,j,i-1}$. This implies that w is a linear combination of vectors that come from subspaces in $\mathcal{A}_{k-1,j,i-1}$, which is a subset of $\mathcal{A}_{k-1,j,i}$. But this contradicts the fact that $\mathcal{A}_{k-1,j,i}$ is a prebasis, as we can write the nonzero vector w both as a linear combination of vectors from subspaces in $\mathcal{A}_{k-1,j,i-1}$ and as a linear combination of vectors from subspaces in $\mathcal{A}_{k-1,j,i} \setminus \mathcal{A}_{k,j,i-1}$, which are

two disjoint subsets of $\mathcal{A}_{k-1,j,i}$. It follows from this contradiction that all the subspaces in $\mathcal{A}_{k,j,i-1} \cup \mathcal{A}_{k-1,j,i}$ are linearly independent of each other. Therefore, $\mathcal{A}_{k,j,i-1} \cup \mathcal{A}_{k-1,j,i}$ is a prebasis for $A_{k,j,i-1} + A_{k-1,j,i}$.

Recall that $A_{kji} \supseteq A_{k,j,i-1} + A_{k-1,j,i}$ and $a_{kji} \in A_{kji} \downarrow (A_{k,j,i-1} + A_{k-1,j,i})$. As $\mathcal{A}_{k,j,i-1} \cup \mathcal{A}_{k-1,j,i}$ is a prebasis for $A_{k,j,i-1} + A_{k-1,j,i}$, $\mathcal{A}_{kji} = \mathcal{A}_{k,j,i-1} \cup \mathcal{A}_{k-1,j,i} \cup \{a_{kji}\}$ is a prebasis for A_{kji}.

A symmetric argument shows that $\mathcal{B}_{kji}$ is a prebasis for B_{kji}. □

Let $\mathcal{A}_j = \mathcal{A}_{Ljj}$; then $\mathcal{A}_j$ is a prebasis for $\mathbb{R}^{d_j}$ (because $A_{Ljj} = \mathbb{R}^{d_j}$) that is a superset of all the other $\mathcal{A}$-prebases for unit layer j. Moreover, $\mathcal{A}_{Lji}$ is a prebasis for $\operatorname{col} W_{j\sim i}$ (because $A_{Lji} = \operatorname{col} W_{j\sim i}$) and $\mathcal{A}_{kjj}$ is a prebasis for $\operatorname{null} W_{k+1\sim j}$ (because $A_{kjj} = \operatorname{null} W_{k+1\sim j}$). So we have found a single prebasis $\mathcal{A}_j$ whose elements simultaneously span many of the subspaces we are interested in!

Similarly, let $\mathcal{B}_j = \mathcal{B}_{jj0}$. Then $\mathcal{B}_j$ is a prebasis for $\mathbb{R}^{d_j}$, $\mathcal{B}_{kj0}$ is a prebasis for $\operatorname{row} W_{k\sim j}$, and $\mathcal{B}_{jji}$ is a prebasis for $\operatorname{null} W_{j\sim i-1}^\top$.

We warn that the prebasis $\mathcal{A}_j$ *cannot*, in general, be chosen so its subspaces are mutually orthogonal. (Nor can $\mathcal{B}_j$.) An orthogonal prebasis is ruled out whenever there is some $\operatorname{null} W_{k+1\sim j}$ and some $\operatorname{col} W_{j\sim i}$ that meet each other at an oblique angle; see A_{410} and A_{311} in Figure 6. Even if $\mathcal{A}_j$ is the standard prebasis (i.e., every subspace we choose from a set of the form $Z \downarrow Y$ is fully orthogonal to Y), we cannot force *all* the prebasis subspaces in $\mathcal{A}_j$ to be mutually orthogonal.

Our prebasis construction permits much flexibility in choosing the prebasis subspaces. But a Fundamental Theorem of Linear Neural Networks seems more satisfying if we explicitly write out the most natural candidates, just as the Fundamental Theorem of Linear Algebra specifies the rowspace, the nullspace, the columnspace, and the left nullspace. Given two subspaces $S, T \in \mathbb{R}^d$, let $\operatorname{proj}_S T$ denote the orthogonal projection of T onto S. Recall our convention that $W_{L+1\sim j} = 0$ and $W_{j\sim -1} = 0$.

Lemma 5. *For $L \geq k \geq j \geq i \geq 0$, the standard prebasis subspaces are*

$$\begin{aligned} a_{kji} &= \operatorname{proj}_{\operatorname{col} W_{j\sim i}} \operatorname{row} W_{k\sim j} \cap \operatorname{proj}_{\operatorname{null} W_{k+1\sim j}} \operatorname{null} W_{j\sim i-1}^\top \\ &= \operatorname{col} W_{j\sim i} \cap (\operatorname{row} W_{k\sim j} + \operatorname{null} W_{j\sim i}^\top) \cap \operatorname{null} W_{k+1\sim j} \cap (\operatorname{row} W_{k+1\sim j} + \operatorname{null} W_{j\sim i-1}^\top) \quad \textit{and} \qquad (4) \\ b_{kji} &= \operatorname{proj}_{\operatorname{row} W_{k\sim j}} \operatorname{col} W_{j\sim i} \cap \operatorname{proj}_{\operatorname{null} W_{j\sim i-1}^\top} \operatorname{null} W_{k+1\sim j} \\ &= \operatorname{row} W_{k\sim j} \cap (\operatorname{null} W_{k\sim j} + \operatorname{col} W_{j\sim i}) \cap \operatorname{null} W_{j\sim i-1}^\top \cap (\operatorname{null} W_{k+1\sim j} + \operatorname{col} W_{j\sim i-1}). \qquad (5) \end{aligned}$$

Proof. In the standard prebasis, from each set of the form $Z \downarrow Y$ we choose the element $Z \cap Y^\perp$. Observe that for two subspaces Z and Y, $Z \cap (Z \cap Y)^\perp = Z \cap (Z^\perp + Y^\perp) = \operatorname{proj}_Z Y^\perp$. Hence

$$\begin{aligned} a_{kji} &= A_{kji} \cap (A_{k,j,i-1} + A_{k-1,j,i})^\perp \\ &= A_{kji} \cap A_{k,j,i-1}^\perp \cap A_{k-1,j,i}^\perp \\ &= \operatorname{null} W_{k+1\sim j} \cap \operatorname{col} W_{j\sim i} \cap (\operatorname{null} W_{k+1\sim j} \cap \operatorname{col} W_{j\sim i-1})^\perp \cap (\operatorname{null} W_{k\sim j} \cap \operatorname{col} W_{j\sim i})^\perp \\ &= \operatorname{proj}_{\operatorname{null} W_{k+1\sim j}} (\operatorname{col} W_{j\sim i-1})^\perp \cap \operatorname{proj}_{\operatorname{col} W_{j\sim i}} (\operatorname{null} W_{k\sim j})^\perp \\ &= \operatorname{proj}_{\operatorname{null} W_{k+1\sim j}} \operatorname{null} W_{j\sim i-1}^\top \cap \operatorname{proj}_{\operatorname{col} W_{j\sim i}} \operatorname{row} W_{k\sim j}. \end{aligned}$$

The third line implies (4). Symmetrically,

$$\begin{aligned}
b_{kji} &= B_{kji} \cap (B_{k,j,i+1} + B_{k+1,j,i})^{\perp} \\
&= B_{kji} \cap B_{k,j,i+1}^{\perp} \cap B_{k+1,j,i}^{\perp} \\
&= \operatorname{row} W_{k\sim j} \cap \operatorname{null} W_{j\sim i-1}^{\top} \cap (\operatorname{row} W_{k\sim j} \cap \operatorname{null} W_{j\sim i}^{\top})^{\perp} \cap (\operatorname{row} W_{k+1\sim j} \cap \operatorname{null} W_{j\sim i-1}^{\top})^{\perp} \\
&= \operatorname{proj}_{\operatorname{row} W_{k\sim j}} (\operatorname{null} W_{j\sim i}^{\top})^{\perp} \cap \operatorname{proj}_{\operatorname{null} W_{j\sim i-1}^{\top}} (\operatorname{row} W_{k+1\sim j})^{\perp} \\
&= \operatorname{proj}_{\operatorname{row} W_{k\sim j}} \operatorname{col} W_{j\sim i} \cap \operatorname{proj}_{\operatorname{null} W_{j\sim i-1}^{\top}} \operatorname{null} W_{k+1\sim j}.
\end{aligned}$$

The third line implies (5). □

4.4 Constructing a Prebasis that Flows through the Network

We have flexibility in choosing a prebasis subspace $a_{kji} \in A_{kji} \downarrow (A_{k,j,i-1} + A_{k-1,j,i})$. Optionally, we can choose *flow prebasis subspaces*: that is, for $j \in [i+1, k]$, we can always choose $a_{kji} = W_j a_{k,j-1,i}$ and $b_{k,j-1,i} = W_j^{\top} b_{kji}$ (after we choose the prebases a_{kii} and b_{kki} with full flexibility; for example, we could choose (4) for a_{kii} and (5) for b_{kki}). These subspaces flow through the linear neural network from specific starting layers to specific stopping layers, thereby outlining how information propagates (or would propagate, if it was there), as expressed by a basis flow diagram such as Figure 5 (top) or Figure 6 (bottom). Lemma 7, below, shows that this construction always yields valid prebases. It also shows that—even if we choose prebases that don't flow (like the standard prebases)—for a fixed i and k, the dimension of a_{kji} is the same for every $j \in [i, k]$.

Lemma 6. *Given that $L \geq k \geq j \geq x \geq i \geq 0$, $W_{j\sim x} a_{kxi}$ has the same dimension as a_{kxi}. Given that $L \geq k \geq y \geq j \geq i \geq 0$, $W_{y\sim j}^{\top} b_{kyi}$ has the same dimension as b_{kyi}.*

Proof. By construction, a_{kxi} is linearly independent of $A_{k-1,x,i} = \operatorname{null} W_{k\sim x} \cap \operatorname{col} W_{x\sim i}$. (That is, $a_{kxi} \cap A_{k-1,x,i} = \{\mathbf{0}\}$.) But $a_{kxi} \subseteq A_{kxi} \subseteq \operatorname{col} W_{x\sim i}$. Hence, every nonzero vector in a_{kxi} is in $\operatorname{col} W_{x\sim i}$ but not in $\operatorname{null} W_{k\sim x} \cap \operatorname{col} W_{x\sim i}$; thus no nonzero vector in a_{kxi} is in $\operatorname{null} W_{k\sim x}$; thus no nonzero vector in a_{kxi} is in $\operatorname{null} W_{j\sim x}$. Therefore, $W_{j\sim x} a_{kxi}$ has the same dimension as a_{kxi}.

A symmetric argument shows that $W_{y\sim j}^{\top} b_{kyi}$ has the same dimension as b_{kyi}. □

Lemma 7 (Basis Flow)**.** *Given that $L \geq k \geq j > x \geq i \geq 0$, $W_{j\sim x} a_{kxi} \in A_{kji} \downarrow (A_{k,j,i-1} + A_{k-1,j,i})$. (Hence, we can choose to set $a_{kji} = W_{j\sim x} a_{kxi}$.) Moreover, every subspace in $A_{kji} \downarrow (A_{k,j,i-1} + A_{k-1,j,i})$ has the same dimension as a_{kxi}.*

Given that $L \geq k \geq y > j \geq i \geq 0$, $W_{y\sim j}^{\top} b_{kyi} \in B_{kji} \downarrow (B_{k,j,i+1} + B_{k+1,j,i})$. (Hence, we can choose to set $b_{kji} = W_{y\sim j}^{\top} b_{kyi}$.) Moreover, every subspace in $B_{kji} \downarrow (B_{k,j,i+1} + B_{k+1,j,i})$ has the same dimension as b_{kyi}.

Proof. By definition, the notation $a_{kji} \in A_{kji} \downarrow (A_{k,j,i-1} + A_{k-1,j,i})$ is equivalent to saying that $A_{kji} = a_{kji} + A_{k,j,i-1} + A_{k-1,j,i}$ and $a_{kji} \cap (A_{k,j,i-1} + A_{k-1,j,i}) = \{\mathbf{0}\}$. We wish to show that $a_{kji} = W_{j\sim x} a_{kxi}$ has both these properties.

To show that $A_{kji} = W_{j\sim x} a_{kxi} + A_{k,j,i-1} + A_{k-1,j,i}$, observe that by Lemma 1, $A_{kji} = W_{j\sim x} A_{kxi}$, $A_{k,j,i-1} = W_{j\sim x} A_{k,x,i-1}$, and $A_{k-1,j,i} = W_{j\sim x} A_{k-1,x,i}$. By assumption, $a_{kxi} \in A_{kxi} \downarrow (A_{k,x,i-1} + A_{k-1,x,i})$, so $A_{kxi} = a_{kxi} + A_{k,x,i-1} + A_{k-1,x,i}$. Pre-multiplying both sides of this identity by $W_{j\sim x}$ confirms that $A_{kji} = W_{j\sim x} a_{kxi} + A_{k,j,i-1} + A_{k-1,j,i}$ (the first property).

To show that $W_{j\sim x} a_{kxi} \cap (A_{k,j,i-1} + A_{k-1,j,i}) = \{\mathbf{0}\}$, let v be a vector in $W_{j\sim x} a_{kxi} \cap (A_{k,j,i-1} + A_{k-1,j,i})$. Then $v \in W_{j\sim x} a_{kxi} \cap W_{j\sim x}(A_{k,x,i-1} + A_{k-1,x,i})$. So there exists a vector $u \in a_{kxi}$ such that $v = W_{j\sim x} u$, and there exist a

vector $s \in A_{k,x,i-1}$ and a vector $t \in A_{k-1,x,i}$ such that $v = W_{j\sim x}(s+t)$. Thus $W_{j\sim x}(u-s-t) = \mathbf{0}$, so $W_{k\sim j}W_{j\sim x}(u-s-t) = \mathbf{0}$ and thus $u - s - t \in \operatorname{null} W_{k\sim x}$. Recall that $A_{k-1,x,i} = \operatorname{null} W_{k\sim x} \cap \operatorname{col} W_{x\sim i}$. So $t \in \operatorname{null} W_{k\sim x}$, hence $u - s \in \operatorname{null} W_{k\sim x}$. Moreover, u and s are both in $\operatorname{col} W_{x\sim i}$, so $u - s \in \operatorname{null} W_{k\sim x} \cap \operatorname{col} W_{x\sim i} = A_{k-1,x,i}$ and hence $u \in A_{k,x,i-1} + A_{k-1,x,i}$. Therefore, $u \in a_{kxi} \cap (A_{k,x,i-1} + A_{k-1,x,i})$. But $a_{kxi} \in A_{kxi} \downarrow (A_{k,x,i-1} + A_{k-1,x,i})$, so $u = \mathbf{0}$ and thus $v = \mathbf{0}$. We have thus shown that every vector in $W_{j\sim x}a_{kxi} \cap (A_{k,j,i-1} + A_{k-1,j,i})$ is $\mathbf{0}$.

Therefore, $W_{j\sim x}a_{kxi} \in A_{kji} \downarrow (A_{k,j,i-1} + A_{k-1,j,i})$, as claimed. To show that every subspace in $A_{kji} \downarrow (A_{k,j,i-1} + A_{k-1,j,i})$ has the same dimension as a_{kxi}, we merely add that $W_{j\sim x}a_{kxi}$ has the same dimension as a_{kxi} by Lemma 6, and the subspaces in $A_{kji} \downarrow (A_{k,j,i-1} + A_{k-1,j,i})$ all have the same dimension as each other.

A symmetric argument shows that $W^{\top}_{y\sim j}b_{kyi} \in B_{kji} \downarrow (B_{k,j,i+1} + B_{k+1,j,i})$ and that every subspace in $B_{kji} \downarrow (B_{k,j,i+1} + B_{k+1,j,i})$ has the same dimension as b_{kyi}. □

While Figure 5 depicts the flow of the prebasis subspaces a_{kji} through the linear neural network, we could depict the transpose flow with nearly the same figure, simply replacing each a_{kji} by b_{kji} and reversing the directions of the arrows. The bottom half of the figure would not change. (To revise Figure 6, we would also need to replace the subspaces depicted with the B_{kji}'s and b_{kji}'s.)

Observe that even if two prebases a_{kji} and $a_{k'ji'}$ at layer j are orthogonal to each other, the prebases $W_{j+1}a_{kji}$ and $W_{j+1}a_{k'ji'}$ generally are not orthogonal. Choosing prebases that "flow" entails sacrificing the desire to choose each layer's prebasis to be as close to orthogonal as possible (i.e., the standard prebasis). But as we have already said, a fully orthogonal prebasis is not generally possible anyway (for example, where a nullspace meets a columnspace obliquely).

4.5 Relationships between Matrix Ranks and Prebasis Subspace Dimensions

This section examines the relationship between the ranks of the subsequence matrices $W_{y\sim x}$ and the dimensions of the prebasis subspaces a_{kji} and b_{kji}. A key insight is that if we know all the subsequence matrix ranks, the dimensions of the prebasis subspaces are uniquely determined, and vice versa (as illustrated at the bottom of Figure 5). To say it another way, there is a bijection between valid rank lists and valid multisets of intervals (with "valid" defined as in Section 4.1).

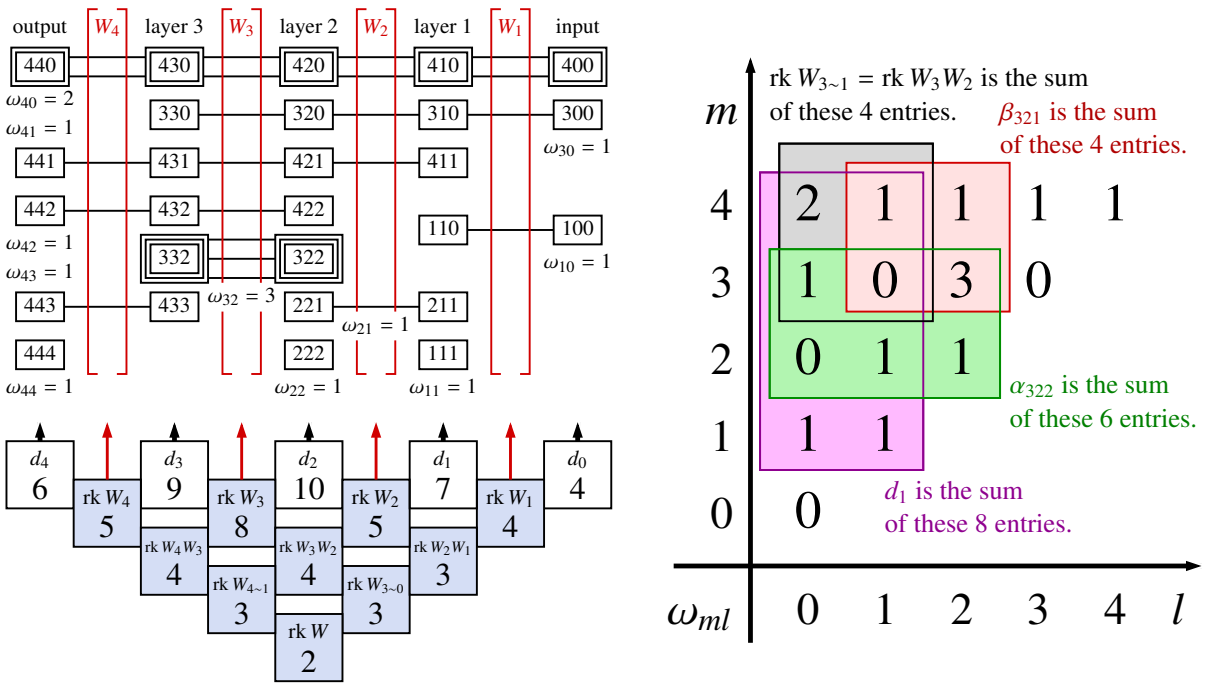
Lemma 7 establishes that the dimension of a_{kji} is the same for every $j \in [i, k]$. By Lemma 3, the dimension of b_{kji} is the same too. So we omit the j indices as we now name this dimension.

Let

$$\omega_{ki} = \dim a_{kji} = \dim b_{kji}, \qquad \text{for all } k, j, i \text{ satisfying } L \geq k \geq j \geq i \geq 0.$$

We have already seen this notation, ω_{ki}, at the start of Section 4, where it denotes the multiplicity of an interval $[i, k]$. Section 4.4 substantiates that connection. The multiplicity ω_{ki} signifies a prebasis subspace a_{kii} of dimension ω_{ki} that originates at layer i, flows through the network being linearly transformed into a sequence of bases $a_{k,i+1,i}, a_{k,i+2,i}, \ldots$, all of dimension ω_{ki}, reaches layer k in the form a_{kki}, and proceeds no farther (either because a_{kki} is in the nullspace of W_{k+1} or because layer k is the output layer), as illustrated in Figures 5 and 6.

The multiplicity ω_{ki} also signifies a prebasis subspace b_{kki} of dimension ω_{ki} that originates at layer k and flows through the transpose network to terminate at layer i in the form b_{kii}. This symmetry surprises us, as sometimes the bases a_{kji} and b_{kji} are necessarily unrelated to each other, except that they have the same dimension. However, the symmetry seems less surprising and even inevitable when you consider that the



$$d_j = \sum_{m=j}^{L}\sum_{l=0}^{j}\omega_{ml} \;\Big|\; \text{rk}\, W_{k\sim i} = \sum_{m=k}^{L}\sum_{l=0}^{i}\omega_{ml} \;\Big|\; \alpha_{kji} = \sum_{m=j}^{k}\sum_{l=0}^{i}\omega_{ml} \;\Big|\; \beta_{kji} = \sum_{m=k}^{L}\sum_{l=i}^{j}\omega_{ml}$$

Figure 7: At left, we reprise the basis flow diagram from Figure 5. At right, we tabulate the values of the interval multiplicities ω_{ml} with boxes that illustrate how the summations compute d_1, α_{322}, β_{321}, and rk $W_{3\sim1}$ = rk W_3W_2. At the bottom of the figure, we reprise the four summations for reference.

fibers $\mu^{-1}(W)$ and $\mu^{-1}(W^\top)$ must be identical (assuming the weight variables are labeled in the right correspondence).

Figure 7 gives a preview of four of the five identities proven in this section—summations that express d_j, rk $W_{k\sim i}$, α_{kji}, and β_{kji} in terms of interval multiplicities ω_{ml}—and a visual interpretation of those summations. The bottom of Figure 5 gives a visual interpretation of the fifth identity, which expresses ω_{ki} in terms of matrix ranks, and a second visual interpretation of the summation for rk $W_{k\sim i}$. It might be helpful to know that the multiplicities ω_{ml} in Figure 7 are the same as in Figure 5, but they are rotated 135°.

Lemma 4 states that $\mathcal{A}_{kji}$ is a prebasis for A_{kji}, where $\mathcal{A}_{kji}$ contains every prebasis subspace $a_{k'ji'}$ with $k' \le k$ and $i' \le i$. The following lemma states that, as we would expect, the dimension of A_{kji} is the sum of the dimensions of the prebases in $\mathcal{A}_{kji}$. But the proof does not directly appeal to Lemma 4; Lemma 3 suffices.

Lemma 8. *For $L \ge k \ge j \ge i \ge 0$, the dimensions α_{kji} of the flow subspaces A_{kji}, the dimensions β_{kji} of the flow subspaces B_{kji}, and the dimensions ω_{ml} of the prebasis subspaces a_{mll} and b_{mll} are related by the identities*

$$\alpha_{kji} = \dim A_{kji} = \sum_{m=j}^{k}\sum_{l=0}^{i}\omega_{ml} \quad \textit{and} \quad \beta_{kji} = \dim B_{kji} = \sum_{m=k}^{L}\sum_{l=i}^{j}\omega_{ml}. \tag{6}$$

Proof. We prove the first claim by induction on increasing values of k and i. For the base cases, recall our convention that $A_{k,j,-1} = \{\mathbf{0}\}$ and $A_{j-1,j,i} = \{\mathbf{0}\}$; hence $\alpha_{k,j,-1} = \alpha_{j-1,j,i} = \alpha_{j-1,j,-1} = 0$.

For the inductive case—the identity for α_{kji}—we assume the inductive hypothesis that the identity holds for $\alpha_{k,j,i-1}$, $\alpha_{k-1,j,i}$, and $\alpha_{k-1,j,i-1}$. By Lemma 3, $\alpha_{kji} = \omega_{ki} + \alpha_{k,j,i-1} + \alpha_{k-1,j,i} - \alpha_{k-1,j,i-1}$. By substituting (6) into the right-hand side, we obtain (6) on the left-hand side, confirming the claim for α_{kji}.

A symmetric argument (by induction on *decreasing* values of k and i), with the identity $\beta_{kji} = \omega_{ki} + \beta_{k,j,i+1} + \beta_{k+1,j,i} - \beta_{k+1,j,i+1}$ from Lemma 3, establishes the identity (6) for β_{kji}. □

The following corollary states that, as we would expect, the number of units d_j in unit layer j equals the sum of the dimensions of the subspaces in a prebasis for $\mathbb{R}^{d_j}$.

Corollary 9. *The number of units in unit layer j is, as formula (1) says,*

$$d_j = \sum_{m=j}^{L} \sum_{l=0}^{j} \omega_{ml}.$$

Proof. As $\mathbb{R}^{d_j} = A_{Ljj} = B_{jj0}$, $d_j = \alpha_{Ljj} = \beta_{jj0}$. The summation follows by identity (6). □

Recall that a rank list is a list of the ranks of all the subsequence matrices (of the form $\operatorname{rk} W_{k\sim i}$), including those of the form $\operatorname{rk} W_{j\sim j} = d_j$, the number of units in unit layer j. The following lemma shows how to map a rank list to a multiset of intervals (expressed as a list of interval multiplicities ω_{ki}) and vice versa. The bottom of Figure 5 depicts the identities (7) and (8).

Lemma 10. *For $L \geq k \geq i \geq 0$, the ranks of the subsequence matrices are related to the dimensions of the flow subspaces and the dimensions of the prebasis subspaces by the identities*

$$\operatorname{rk} W_{k\sim i} = \alpha_{Lki} = \beta_{ki0} = \sum_{m=k}^{L} \sum_{l=0}^{i} \omega_{ml} \quad \textit{and} \tag{7}$$

$$\omega_{ki} = \operatorname{rk} W_{k\sim i} - \operatorname{rk} W_{k\sim i-1} - \operatorname{rk} W_{k+1\sim i} + \operatorname{rk} W_{k+1\sim i-1}, \tag{8}$$

recalling the conventions that $\operatorname{rk} W_{j\sim j} = d_j$ *and* $\operatorname{rk} W_{L+1\sim x} = 0 = \operatorname{rk} W_{y\sim -1}$.

Proof. We use the Rank-Nullity Theorem to connect the rank of $W_{k\sim i}$ to the dimensions of the flow subspaces, and the formulae (6) to connect those to the interval multiplicities. Recall that $A_{Lii} = \mathbb{R}^{d_i}$ and $A_{k-1,i,i} = \operatorname{null} W_{k\sim i} \cap \operatorname{col} W_{i\sim i} = \operatorname{null} W_{k\sim i}$. As $W_{k\sim i}$ is a $d_k \times d_i$ matrix,

$$\begin{aligned}
\operatorname{rk} W_{k\sim i} &= d_i - \dim \operatorname{null} W_{k\sim i} \\
&= \dim A_{Lii} - \dim A_{k-1,i,i} \\
&= \alpha_{Lii} - \alpha_{k-1,i,i} \\
&= \sum_{m=k}^{L} \sum_{l=0}^{i} \omega_{ml} \\
&= \alpha_{Lki} = \beta_{ki0}
\end{aligned}$$

as claimed. (Symmetrically, we could obtain the summation (7) by instead starting from $\operatorname{rk} W_{k\sim i} = d_k - \dim \operatorname{null} W_{k\sim i}^{\top}$ and recalling that $B_{kk0} = \mathbb{R}^{d_k}$ and $B_{k,k,i+1} = \operatorname{null} W_{k\sim i}^{\top}$. This is how we originally realized that $\dim a_{kji} = \dim b_{kji}$, which led us to Lemma 2.)

We can verify the identity $\operatorname{rk} W_{k\sim i} - \operatorname{rk} W_{k\sim i-1} - \operatorname{rk} W_{k+1\sim i} + \operatorname{rk} W_{k+1\sim i-1} = \omega_{ki}$ by substituting the summation (7) into it. □

Ferdinand Georg Frobenius [5] proved in 1911 that $\operatorname{rk} W_{k\sim i} - \operatorname{rk} W_{k\sim i-1} - \operatorname{rk} W_{k+1\sim i} + \operatorname{rk} W_{k+1\sim i-1} \geq 0$, a statement called the *Frobenius rank inequality*. This confirms that every ω_{ml} is nonnegative. Our derivations deepen the Frobenius rank inequality by connecting the slack (8) in the inequality to the dimension of the subspaces (4) and (5). (We considered calling each interval multiplicity ω_{ml} a *Frobenius slack*.) To put it in a simpler notation,

$$\begin{aligned} \operatorname{rk} ST - \operatorname{rk} STU - \operatorname{rk} RST + \operatorname{rk} RSTU &= \dim(\operatorname{proj}_{\operatorname{col} T} \operatorname{row} S \cap \operatorname{proj}_{\operatorname{null} RS} \operatorname{null}(TU)^{\top}) \\ &= \dim(\operatorname{proj}_{\operatorname{row} S} \operatorname{col} T \cap \operatorname{proj}_{\operatorname{null}(TU)^{\top}} \operatorname{null} RS). \end{aligned}$$

Thome [15] offers some generalizations of the Frobenius rank inequality to linear neural networks. He proves them by induction, but they also follow easily from (7).

4.6 A Fundamental Theorem of Linear Neural Networks?

Matrices have a few crucial properties that Gilbert Strang [14] summarizes as a *Fundamental Theorem of Linear Algebra*. For any $y \times z$ matrix W, the rowspace of W (denoted $\operatorname{row} W$) is orthogonal to the nullspace of W (denoted $\operatorname{null} W$) and the sum of their dimensions is z (the latter fact is known as the *Rank-Nullity Theorem*). The same observation applies to $W^{\top}$, so the columnspace of W (denoted $\operatorname{col} W$) is orthogonal to the left nullspace of W (denoted $\operatorname{null} W^{\top}$) and the sum of their dimensions is y. The dimensions of $\operatorname{row} W$ and $\operatorname{col} W$ are the same (namely $\operatorname{rk} W$). (Strang's presentation also includes two properties of the singular value decomposition, not treated here.)

Here, we outline a candidate for an analogous *Fundamental Theorem of Linear Neural Networks*. That candidate is the combination of Lemmas 3, 4, 6, and 7, and the formulae (1), (2), (6), (7), and (8), with the standard basis subspaces (4) and (5) giving us canonical examples of a decomposition of each layer of units into linearly independent subspaces. If we replace the standard prebasis with a flow prebasis, our Fundamental Theorem gives us additional insight into the flow of information through a linear neural network.

These results do not lend themselves to a pithy statement like the Fundamental Theorem of Linear Algebra, but they are nevertheless useful, as our forthcoming work will demonstrate. This paper was motivated by our studies of the topology and geometry of the fiber $\mu^{-1}(W)$ of a matrix W, where the flow prebases play a large role in showing how manifolds of different dimensions are knitted together to form the fiber, and in enumerating the ways we can locally modify the weights of a linear neural network without changing the linear transformation that the network computes. These insights, in turn, have applications to understanding critical points in the cost functions used to train neural networks [16], because gradient descent algorithms sometimes make progress by linking subspace flows together.

5 Determinantal Manifolds, Metadeterminantal Manifolds, and Strata

The fiber $\mu^{-1}(W)$ of a matrix W is not generally a manifold, but we claim that we can understand the fiber by partitioning it into a finite number of disjoint strata, which are manifolds of different dimensions. Each stratum corresponds to a different rank list—or equivalently, to a different multiset of intervals. Before we partition the fiber into strata, it is useful to consider how the entire weight space $\mathbb{R}^{d_\theta}$ can be partitioned into a set of *metadeterminantal manifolds*, one for each possible rank list.

We define the *rank list* $\underline{r}$ for a weight vector $\theta = (W_L, W_{L-1}, \ldots, W_1) \in \mathbb{R}^{d_\theta}$ to be a sequence specifying the ranks of all the subsequence matrices, i.e., $\underline{r} = \langle \operatorname{rk} W_{k\sim i} \rangle_{L \geq k \geq i \geq 0}$. The rank list includes the unit layer sizes $\operatorname{rk} W_{j\sim j} = d_j$, and we assume these are fixed.

Sometimes we do not want to specify a particular θ, but rather we wish to specify some target ranks; in this case we let $r_{k\sim i}$ denote the target value of $\operatorname{rk} W_{k\sim i}$ and we write $\underline{r} = \langle r_{k\sim i} \rangle_{L \geq k \geq i \geq 0}$. Consider the set of all points in weight space whose subsequence matrices all match a specified rank list $\underline{r}$. We call this set a *metadeterminantal manifold*, which we denote

$$MM_{\underline{r}} = \{\theta = (W_L, \ldots, W_1) \in \mathbb{R}^{d_\theta} : \operatorname{rk} W_{k\sim i} = r_{k\sim i} \text{ for all } L \geq k > i \geq 0\}.$$

To understand $MM_{\underline{r}}$, it helps to examine one matrix at a time. There is a well known algebraic variety called the *determinantal variety*, which we denote

$$DV_r^{y\times z} = \{M \in \mathbb{R}^{y\times z} : \operatorname{rk} M \leq r\},$$

the set of all $y \times z$ real matrices of rank at most r. The determinantal variety has singular points and thus is not a manifold (unless the rank is zero; $DV_0^{y\times z}$ contains only the zero matrix). It is well known that the singular locus of $DV_r^{y\times z}$ is $DV_{r-1}^{y\times z}$. If we omit matrices of rank strictly less than r, we obtain a manifold that we call the *determinantal manifold*, denoted

$$DM_r^{y\times z} = DV_r^{y\times z} \setminus DV_{r-1}^{y\times z} = \{M \in \mathbb{R}^{y\times z} : \operatorname{rk} M = r\}.$$

Observe that $DV_r^{y\times z}$ is the closure of $DM_r^{y\times z}$. So although $DM_r^{y\times z}$ is a manifold, it is not closed with respect to the weight space. It is well known that both $DM_r^{y\times z}$ and $DV_r^{y\times z}$ have dimension $r(y + z - r)$.

If a weight vector $\theta = (W_L, \ldots, W_1)$ lies on the metadeterminantal manifold $MM_{\underline{r}}$, then W_1 must lie on the determinantal manifold $DM_{r_{1\sim 0}}$, W_2 must lie on the determinantal manifold $DM_{r_{2\sim 1}}$, and so on. But we also have to constrain the ranks of subsequence matrices like $W_{4\sim 0}$ that are not factor matrices. The restriction that θ must satisfy $\operatorname{rk} W_{4\sim 0} = r_{4\sim 0}$ motivates *weight-space determinantal manifolds*, denoted

$$WDM_{\underline{r}}^{k\sim i} = \{\theta \in \mathbb{R}^{d_\theta} : \operatorname{rk} W_{k\sim i} = r_{k\sim i}\}.$$

Unfortunately, these are not as well studied as ordinary determinantal manifolds. Each weight-space determinantal manifold is an algebraic variety, as it is the set of solutions of a system of polynomial equations. These equations are found by setting all the order-$(r_{k\sim i} + 1)$ minors of $W_{k\sim i}$ to zero. It is not obvious that each $WDM_{\underline{r}}^{k\sim i}$ is a manifold.

A second way to define the metadeterminantal manifold is as the intersection of the weight-space determinantal manifolds:

$$MM_{\underline{r}} = \bigcap_{L \geq k > i \geq 0} WDM_{\underline{r}}^{k\sim i}.$$

Every point θ in weight space lies on one metadeterminantal manifold (the one with the correct rank list for θ), so the metadeterminantal manifolds partition the entire weight space $\mathbb{R}^{d_\theta}$ into manifolds of various dimensions. Given a matrix W, we can stratify its fiber $\mu^{-1}(W)$ by creating, for each rank list $\underline{r}$ that is valid for W, a stratum

$$S_{\underline{r}}^W = \mu^{-1}(W) \cap MM_{\underline{r}}.$$

That is, $S_{\underline{r}}^W$ is the set of points $\theta \in \mathbb{R}^{d_\theta}$ such that $\mu(\theta) = W$ and $\underline{r}$ is the rank list for θ. We claim that $S_{\underline{r}}^W$ is a manifold. In the next section, we derive its dimension and its tangent space at θ.

Note that $S_{\underline{r}}^W$ is empty if $r_{L\sim 0} \neq \operatorname{rk} W$. We only need to consider rank lists that get the rank of W right.

6 Moves on and off the Fiber

Imagine you are standing at a point θ on a fiber $\mu^{-1}(W)$. A *move* (θ, θ') is a step you take from θ to another point θ', which may or may not be on the fiber. Let $\Delta\theta = \theta' - \theta$ be the *displacement* of the move. We write

$$\begin{aligned} \theta' &= (W'_L, W'_{L-1}, \ldots, W'_1) \in \mathbb{R}^{d_\theta} \quad \text{and} \\ \Delta\theta &= (\Delta W_L, \Delta W_{L-1}, \ldots, \Delta W_1) \in \mathbb{R}^{d_\theta}. \end{aligned}$$

We use analogous notation for the product $W' = \mu(\theta')$, its displacement $\Delta W = W' - W$, the modified subsequence matrices $W'_{j\sim i} = W'_j W'_{j-1} \cdots W'_{i+1}$, and their displacements $\Delta W_{j\sim i} = W'_{j\sim i} - W_{j\sim i}$.

Not every displacement constitutes a "move." As we define it, a move requires $\Delta\theta$ to lie in one of the subspaces we will define in Section 6.1 and 6.5. (Even this statement has to be modified to account for the curvature of the fiber also complicates the definition of "move.") Most of these subspaces are tangent to one or more strata that adjoin the point θ. They will help us to count the dimensions of strata and geometrically characterize their connections to other strata.

Two classes of move suffice to characterize strata, their tangent spaces, and their interconnections: one-matrix moves and two-matrix moves.

- A *one-matrix move* has at most one nonzero displacement matrix ΔW_j. That is, $W'_z = W_z$ for all $z \neq j$. Moreover, we require that $\operatorname{col} \Delta W_j \subseteq a_{lji}$ for some prebasis subspace a_{lji} (defined in Section 4.3) with $L \geq l \geq j \geq i \geq 0$, and that $\operatorname{row} \Delta W_j \subseteq b_{k,j-1,h}$ for some prebasis subspace $b_{k,j-1,h}$ with $L \geq k \geq j-1 \geq h \geq 0$. In a more concise notation, $\Delta W_j \in a_{lji} \otimes b_{k,j-1,h}$; we call $a_{lji} \otimes b_{k,j-1,h}$ a *one-matrix subspace*. Some one-matrix moves stay on the fiber (specifically, we will see that $W' = W$ if $L > l$ or $i > 0$) and some move off of it ($W' \neq W$ if $l = L$ and $i = 0$, unless $\Delta W_j = 0$). Some of the one-matrix moves that stay on the fiber also stay on the stratum that contains θ (that is, no subsequence matrix changes its rank), and some move off of it (some subsequence matrix changes its rank).

  One-matrix moves are *linear* in two senses. First, the displacement ΔW is linear in the displacement ΔW_j. Second, as a consequence, if a one-matrix move stays on the fiber ($W' = W$), then for all $\kappa \in \mathbb{R}$, $\mu(\theta + \kappa\Delta\theta) = W$. That is, the line through θ and θ' is a subset of the fiber. Moreover, if you have a set of one-matrix displacements that all displace W_j and all stay on the fiber, then any linear combination of those displacements is a displacement that also stays on the fiber. Section 6.1 discusses one-matrix moves in detail.

- A *two-matrix move* has exactly two nonzero displacement matrices ΔW_{j+1} and ΔW_j (which are always consecutive). In a *finite two-matrix move*, we choose a matrix $K \in a_{lji} \otimes b_{kjh}$ for some a_{lji} with $L \geq l > j \geq i \geq 0$ and some b_{kjh} with $L \geq k \geq j > h \geq 0$, then set $W'_{j+1} = W_{j+1}(I + K)$ and $W'_j = (I + K)^{-1} W_j$. (We assume that $I + K$ is invertible; it always is if K is sufficiently small.) It is easy to see that *every* finite two-matrix move stays on the fiber ($W' = W$) and moreover stays on the same stratum as θ (no subsequence matrix changes its rank).

  Unlike in a one-matrix move, often there is no straight path on the fiber from θ to θ'. But there is a natural choice of a curved path that stays on the fiber, and moreover stays on the same stratum as θ. Of particular interest to us is the initial direction of motion from θ as you walk along that path—that is, the direction tangent to the path at θ. That direction is the same as the direction of the displacement determined by $\Delta W_{j+1} = W_{j+1}K$ and $\Delta W_j = -KW_j$ (ignoring its magnitude). We call a displacement in that direction a *differential two-matrix move*. A move in that direction doesn't stay on the fiber (because of its curvature), but it stays on a line tangent to the stratum containing θ. By characterizing

all differential two-matrix moves, we can determine the stratum's tangent space and the dimension of the stratum. Section 6.5 discusses two-matrix moves in detail.

We will consider both *finite moves*, which are simple moves from one point to another, and *infinitesimal moves*: moves in the limit as $\Delta\theta$ becomes infinitesimally small.

Infinitesimal moves are motivated by two considerations. First, an infinitesimal perturbation of a matrix can increase its rank, but cannot decrease its rank—decreasing the rank entails a finite displacement. By studying moves that are so small that no subsequence matrix can decrease in rank, we simplify understanding how strata of different dimensions are connected to each other. For example, in Figure 3, the 0-dimensional stratum S_{000} lies in the closure of the 1-dimensional stratum S_{010}, and both of those lie in the closure of the 2-dimensional stratum S_{011}. Starting from any point $\theta \in S_{010}$, an infinitesimal move can reach some point $\theta' \in S_{011}$, which entails an increase in the rank of W_1 from 0 to 1. But from θ, an infinitesimal move does not suffice to reach S_{000}, as θ is not in the closure of S_{000}. Recall that at θ, $W_2 = [\theta_2]$, but at S_{000}, $\operatorname{rk} W_2 = 0$; so the distance from θ to S_{000} is $|\theta_2|$. In general, decreasing the rank of any subsequence matrix always entails moving some finite distance.

Infinitesimals have a complicated status in the history of mathematical rigor. To strip away everything that is not essential, we now define an *infinitesimal move* to be any move such that every stratum whose closure contains θ' also contains θ. That is, you can never enter a stratum's closure by an infinitesimal move if you're not already there. This definition has a counterintuitive consequence: if an infinitesimal move increases the rank of some subsequence matrix, then the inverse move is *not* infinitesimal. Moving from S_{010} to S_{011} is infinitesimal, but moving back is not. If an infinitesimal move moves from one stratum to a different one, the latter stratum has higher dimension than the former.

The second motivation is that most fibers have some curvature. In the limit as a displacement $\Delta\theta$ approaches zero (while keeping θ' on the fiber), $\Delta\theta$ becomes arbitrarily close to tangent to some smooth path on the fiber adjoining θ. Roughly speaking, we want to characterize the subspace tangent to the fiber at θ. Unfortunately, the most interesting points on the fiber are the singular points where the fiber is not locally a manifold and a tangent space is not defined! Fortunately, our stratification partitions the fiber into manifolds (of different dimensions), and we can characterize the tangent space of each stratum whose closure contains θ. All of these tangent spaces are subspaces of a particular subspace, the nullspace of the differential map of $\mu(\theta)$ (see Section 6.4). We use a combination of infinitesimal one-matrix moves and differential two-matrix moves to build prebases that span these subspaces.

In this section, we construct prebases in weight space that can express moves that are aligned with the fiber and each stratum at θ. We identify which infinitesimal one-matrix stay on the same stratum as θ, and which ones move to another stratum. We will construct three different prebases.

- The *one-matrix move prebasis* is a prebasis that spans the entire weight space $\mathbb{R}^{d_\theta}$. Each member of the prebasis is a subspace of one-matrix moves.

- The *fiber prebasis* is a prebasis that spans the nullspace of the differential map of $\mu(\theta)$. Unlike the one-matrix move prebasis, the fiber prebasis does not span the entire weight space $\mathbb{R}^{d_\theta}$. However, the nullspace of the differential map includes all lines tangent at θ to a smooth path on the fiber. The fiber prebasis contains all the one-matrix subspaces whose moves stay on the fiber, excludes all the one-matrix subspaces whose moves do not, and adds some two-matrix moves that represent directions tangent to curved paths on the fiber.

  The fiber prebasis gives us a practical blueprint for moving on the fiber. For example, one could modify a linear neural networks' weights so that it computes the same function as before but it is no longer close to a spurious critical point, thereby helping gradient descent to run faster.

- Recall that our stratification partitions the fiber $\mu^{-1}(W)$ into strata; let S be the stratum that contains θ. The *stratum prebasis* is a prebasis that spans $T_\theta S$, the tangent space of S at θ. The fiber prebasis contains all the one-matrix subspaces whose moves stay on the stratum and adds more two-matrix moves than the fiber prebasis does.

Some one-matrix moves change the ranks of one or more subsequence matrices, thereby moving from one stratum to another; we call them *combinatorial moves*. They are particularly interesting, and we study them in Sections 6.2 and 6.3. There are no combinatorial two-matrix moves; two-matrix moves do not change the rank of any subsequence matrix. We focus primarily on infinitesimal combinatorial moves, which increase the rank of at least one subsequence matrix, but cannot decrease any ranks.

A combinatorial move implies that θ' has a different rank list and a different multiset of intervals than θ, so θ' does not lie in the same stratum as θ. An infinitesimal combinatorial move further implies that θ' lies in a higher-dimensional stratum that has θ on its relative boundary. We subdivide combinatorial moves into two categories: *connecting moves*, which connect two intervals together by increasing the rank of a factor matrix W_j; and *swapping moves*, which replace two intervals by two different intervals, one longer than both the replaced intervals, and one shorter than both. A swapping move does *not* increase the rank of any factor matrix W_j, but both types of moves increase the ranks of one or more subsequence matrices.

6.1 One-Matrix Moves and the One-Matrix Move Prebasis

In a one-matrix move, we choose one finite displacement ΔW_j and set $\Delta W_z = 0$ for all $z \neq j$. (We permit ΔW_j to be zero as well, so our moves include a "move" that doesn't move.) Thus, we move from a point $\theta \in \mu^{-1}(W)$ to

$$\theta' = (W_L, \ldots, W_{j+1}, W'_j, W_{j-1}, \ldots, W_1)$$

where $W'_j = W_j + \Delta W_j$. Then $W' = \mu(\theta') = \mu(\theta) + W_L \cdots W_{j+1} \Delta W_j W_{j-1} \cdots W_1 = W + W_{L\sim j} \Delta W_j W_{j-1\sim 0}$. A displacement ΔW_j has the property that θ' lies on the fiber $\mu^{-1}(W)$ if and only if $W_{L\sim j} \Delta W_j W_{j-1\sim 0} = 0$. The set of displacements that have this property is the subspace

$$\begin{aligned} N_j &= \operatorname{null} W_{L\sim j} \otimes \mathbb{R}^{d_{j-1}} + \mathbb{R}^{d_j} \otimes \operatorname{null} W_{j-1\sim 0}^\top \\ &= A_{L-1,j,j} \otimes B_{j-1,j-1,0} + A_{Ljj} \otimes B_{j-1,j-1,1}. \end{aligned}$$

Here, the symbol "$\otimes$" denotes a tensor product. For linear subspaces $U \subseteq \mathbb{R}^y$ and $V \subseteq \mathbb{R}^x$,

$$U \otimes V = \{M \in \mathbb{R}^{y \times x} : \operatorname{col} M \subseteq U \text{ and } \operatorname{row} M \subseteq V\}.$$

That is, $U \otimes V$ is the set containing every $y \times x$ matrix M such that M maps all points in $\mathbb{R}^x$ into U and $M^\top$ maps all points in $\mathbb{R}^y$ into V.

Observe that $N_j \subseteq \mathbb{R}^{d_j \times d_{j-1}}$ and its dimension is

$$\begin{aligned} \dim N_j &= \dim(\operatorname{null} W_{L\sim j} \otimes \mathbb{R}^{d_{j-1}}) + \dim(\mathbb{R}^{d_j} \otimes \operatorname{null} W_{j-1\sim 0}^\top) - \dim(\operatorname{null} W_{L\sim j} \otimes \operatorname{null} W_{j-1\sim 0}^\top) \\ &= (d_j - \operatorname{rk} W_{L\sim j}) \cdot d_{j-1} + d_j \cdot (d_{j-1} - \operatorname{rk} W_{j-1\sim 0}) - (d_j - \operatorname{rk} W_{L\sim j}) \cdot (d_{j-1} - \operatorname{rk} W_{j-1\sim 0}) \\ &= d_j d_{j-1} - \operatorname{rk} W_{L\sim j} \cdot \operatorname{rk} W_{j-1\sim 0}. \end{aligned}$$

Recall from Section 4.3 that we decompose each unit layer's space $\mathbb{R}^{d_j}$ into a *prebasis*—a "basis" made up of subspaces—which could easily be further decomposed into a basis of vectors (a more familiar concept).

Here we apply the same idea to the factor matrix spaces $\mathbb{R}^{d_j \times d_{j-1}}$ and the weight space $\mathbb{R}^{d_\theta}$. Above, we have expressed N_j in terms of some flow subspaces A_{ljj} and $B_{j-1,j-1,h}$ defined in Section 4.2. This gives us a hint about how we might decompose $\mathbb{R}^{d_\theta}$ into a prebasis that separates moves that stay on the fiber from those that do not.

First, we construct a prebasis $\mathcal{O}_j$ for $\mathbb{R}^{d_j \times d_{j-1}}$. For indices l, k, j, i, and h satisfying $L \geq l \geq j \geq i \geq 0$ and $L \geq k \geq j-1 \geq h \geq 0$, define the prebasis subspace

$$o_{lkjih} = a_{lji} \otimes b_{k,j-1,h},$$

where a_{lji} and $b_{k,j-1,h}$ are the prebasis subspaces defined in Section 4.3. Observe that $\dim o_{lkjih} = \dim a_{lji} \cdot \dim b_{k,j-1,h} = \omega_{li}\,\omega_{kh}$. For each $j \in [1, L]$, define the prebasis

$$\mathcal{O}_j = \{o_{lkjih} \neq \{0\} : l \in [j, L], k \in [j-1, L], i \in [0, j], h \in [0, j-1]\}.$$

This prebasis pairs every subspace in the prebasis $\mathcal{A}_j$ with every subspace in the prebasis $\mathcal{B}_{j-1}$.

Lemma 11. *$\mathcal{O}_j$ is a prebasis for $\mathbb{R}^{d_j \times d_{j-1}}$. In particular, the subspaces in $\mathcal{O}_j$ are linearly independent.*

Proof. □

Note that is is easy to find a basis for o_{lkjih} as follows. Let the vectors $u_1, \ldots, u_\omega \in \mathbb{R}^{d_j}$ be a basis for a_{lji} (where $\omega = \omega_{li}$ is the number of basis vectors) and let $v_1, \ldots, v_{\omega'} \in \mathbb{R}^{d_{j-1}}$ be a basis for $b_{k,j-1,h}$ (where $\omega' = \omega_{kh}$). Then a basis for o_{lkjih} is the set $\{u_i v_j^\top : i \in [1, \omega], j \in [1, \omega']\}$, which contains $\omega_{li}\,\omega_{kh}$ rank-1 outer product matrices. If we take the union of these bases over every $o_{lkjih} \in \mathcal{O}_j$, we have a total of $d_j d_{j-1}$ basis "vectors" (matrices) that form a basis for $\mathbb{R}^{d_j \times d_{j-1}}$.

Now we construct a prebasis Θ_{O} for $\mathbb{R}^{d_\theta}$ that we call the *one-matrix move prebasis*. The subspaces in Θ_{O} have the form

$$\phi_{lkjih} = \{(0, \ldots, 0, M, 0, \ldots, 0) : M \in o_{lkjih}\}$$

with M in position j from the right. Let

$$\Theta_{\mathrm{O}} = \{\phi_{lkjih} \neq \{\mathbf{0}\} : L \geq l \geq j \geq i \geq 0 \text{ and } L \geq k \geq j-1 \geq h \geq 0\}.$$

It is clear that Θ_{O} is a prebasis for $\mathbb{R}^{d_\theta}$, as each $\mathcal{O}_j$ is a prebasis for $\mathbb{R}^{d_j \times d_{j-1}}$.

We are finally ready to define a *one-matrix move*: it is a move with displacement $\Delta\theta \in \phi_{lkjih}$ for some subspace $\phi_{lkjih} \in \Theta_{\mathrm{O}}$.

Next, we distinguish one-matrix moves that stay on the fiber from those that move off the fiber. This motivates the following subsets of $\mathcal{O}_j$ and Θ_{O}.

$$\begin{aligned}
\mathcal{O}_j^{L0} &= \{o_{lkjih} \in \mathcal{O}_j : l = L \text{ and } h = 0\} = \{o_{Lkji0} \neq \{0\} : k \in [j-1, L], i \in [0, j]\}, \\
\Theta_{\mathrm{O}}^{L0} &= \{\phi_{lkjih} \in \Theta_{\mathrm{O}} : l = L \text{ and } h = 0\} = \{\phi_{Lkji0} \neq \{\mathbf{0}\} : L \geq k \geq j-1 \geq 0 \text{ and } L \geq j \geq i \geq 0\}, \\
\mathcal{O}_j^{\text{fiber}} &= \mathcal{O}_j \setminus \mathcal{O}_j^{L0} = \{o_{lkjih} \in \mathcal{O}_j : L > l \text{ or } h > 0\}, \quad \text{and} \\
\Theta_{\mathrm{O}}^{\text{fiber}} &= \Theta_{\mathrm{O}} \setminus \Theta_{\mathrm{O}}^{L0} = \{\phi_{lkjih} \in \Theta_{\mathrm{O}} : L > l \text{ or } h > 0\}.
\end{aligned}$$

We claim that $\mathcal{O}_j^{\text{fiber}}$ is a prebasis for N_j, and that every displacement $\Delta\theta \in \phi_{lkjih}$ with $\phi_{lkjih} \in \Theta_{\mathrm{O}}^{\text{fiber}}$ stays on the fiber; that is, $\mu(\theta + \Delta\theta) = W$. We also claim that every displacement $\Delta\theta \in \phi_{lkjih} \setminus \{\mathbf{0}\}$ with $\phi_{lkjih} \in \Theta_{\mathrm{O}}^{L0}$ leaves the fiber; that is, $\mu(\theta + \Delta\theta) \neq W$.

Lemma 12. *O_j^{fiber} is a prebasis for N_j.*

Proof. Recall that $N_j = A_{L-1,j,j} \otimes \mathbb{R}^{d_{j-1}} + \mathbb{R}^{d_j} \otimes B_{j-1,j-1,1}$. By Lemma 4,

$$\begin{aligned}
\mathbb{R}^{d_j} = A_{Ljj} &= \sum_{l=j}^{L} \sum_{i=0}^{j} a_{lji}, \\
A_{L-1,j,j} &= \sum_{l=j}^{L-1} \sum_{i=0}^{j} a_{lji}, \\
\mathbb{R}^{d_{j-1}} = B_{j-1,j-1,0} &= \sum_{k=j-1}^{L} \sum_{h=0}^{j-1} b_{k,j-1,h}, \qquad \text{and} \\
B_{j-1,j-1,1} &= \sum_{k=j-1}^{L} \sum_{h=1}^{j-1} b_{k,j-1,h}.
\end{aligned}$$

Therefore,

$$\begin{aligned}
A_{L-1,j,j} \otimes \mathbb{R}^{d_{j-1}} &= \sum_{l=j}^{L-1} \sum_{i=0}^{j} \sum_{k=j-1}^{L} \sum_{h=0}^{j-1} a_{lji} \otimes b_{k,j-1,h} = \operatorname{span}\{o_{lkjih} \in O_j : L > l\}, \\
\mathbb{R}^{d_j} \otimes B_{j-1,j-1,1} &= \sum_{l=j}^{L} \sum_{i=0}^{j} \sum_{k=j-1}^{L} \sum_{h=1}^{j-1} a_{lji} \otimes b_{k,j-1,h} = \operatorname{span}\{o_{lkjih} \in O_j : h > 0\}, \qquad \text{and} \\
N_j &= A_{L-1,j,j} \otimes \mathbb{R}^{d_{j-1}} + \mathbb{R}^{d_j} \otimes B_{j-1,j-1,1} = \operatorname{span} O_j^{\text{fiber}}.
\end{aligned}$$

By Lemma 12, the subspaces in O_j are linearly independent; hence so are the subspaces in O_j^{fiber}. Therefore, O_j^{fiber} is a prebasis for N_j. □

The following corollary shows that $\Theta_{\text{O}}^{\text{fiber}}$ represents the one-matrix moves that stay on the fiber, whereas Θ_{O}^{L0} represents the one-matrix moves that move off the fiber (plus the move with $\Delta\theta = 0$, as every subspace must include the trivial move).

Corollary 13. *For every subspace $\phi_{lkjih} \in \Theta_{\text{O}}^{\text{fiber}}$ and every displacement $\Delta\theta \in \phi_{lkjih}$, $\mu(\theta + \Delta\theta) = W$. For every subspace $\phi_{lkjih} \in \Theta_{\text{O}}^{L0}$ and every displacement $\Delta\theta \in \phi_{lkjih} \setminus \{\mathbf{0}\}$, $\mu(\theta + \Delta\theta) \neq W$.*

Proof. Every subspace $\phi_{lkjih} \in \Theta_{\text{O}}$ is a one-matrix move subspace, so a displacement $\Delta\theta \in \phi_{lkjih}$ has at most one nonzero matrix, $\Delta W_j \in o_{lkjih}$. Let $\theta' = \theta + \Delta\theta$, and recall that $\mu(\theta') = W + W_{L\sim j}\Delta W_j W_{j-1\sim 0}$. If $\phi_{lkjih} \in \Theta_{\text{O}}^{\text{fiber}}$, then $\Delta W_j \in N_j$ by Lemma 12, so $\mu(\theta + \Delta\theta) = W$. If $\phi_{lkjih} \in \Theta_{\text{O}}^{L0}$ and $\Delta\theta \in \phi_{lkjih} \setminus \{\mathbf{0}\}$, then $\Delta W_j \notin N_j$ by Lemma 12, so $\mu(\theta + \Delta\theta) \neq W$. □

6.2 The Effects of One-Matrix Moves

There is a crucial distinction between one-matrix moves that change the rank of some subsequence matrix—the combinatorial moves—and one-matrix moves that do not. Following a combinatorial move, θ' is in a different stratum than θ (usually of a different dimension), θ' has a different rank list than θ, θ' has a different multiset of intervals than θ, and usually (but not always) the number of degrees of freedom of motion along the fiber is different at θ' than at θ.

For each subspace o_{lkjih} in the prebasis $\mathcal{O}_j$, we ask: which subsequence matrices change when we replace W_j with $W'_j = W_j + \Delta W_j$, where $\Delta W_j \in o_{lkjih}$? Which subsequence matrices undergo a change in rowspace or columnspace? Which subsequence matrices change rank? The rest of this section answers these questions. Table 4 summarizes the answers for infinitesimal moves.

Let $\Delta W_j = \epsilon p q^\top$ for a scalar $\epsilon \in \mathbb{R}$ and two vectors $p \in a_{lji} \setminus \{\mathbf{0}\}$ and $q \in b_{k,j-1,h} \setminus \{\mathbf{0}\}$. Then $\Delta W_j \in o_{lkjih}$. We assume $l \geq j \geq i$ and $k \geq j-1 \geq h$ (otherwise a_{lji} or $b_{k,j-1,h}$ is not defined). Let $W'_j = W_j + \Delta W_j$, let $\theta = (W_L, W_{L-1}, \ldots, W_1)$, and let θ' be θ with W_j replaced by W'_j. For each subsequence matrix $W_{y\sim x}$, let $W'_{y\sim x}$ denote its new value for θ', and let $\Delta W_{y\sim x} = W'_{y\sim x} - W_{y\sim x}$. The following lemma identifies which subsequence matrices do or do not change.

Lemma 14. *Given $L \geq y \geq x \geq 0$, $W'_{y\sim x} = W_{y\sim x}$ if and only if $\epsilon = 0$ or $j \notin [x+1, y]$ or $y > l$ or $x < h$.*

Proof. If $j \notin [x+1, y]$, then W'_j is not one of the matrices constituting $W'_{y\sim x}$, so $W'_{y\sim x} = W_{y\sim x}$ as claimed.

Otherwise, $\Delta W_{y\sim x} = W_{y\sim j} \Delta W_j W_{j-1\sim x} = \epsilon W_{y\sim j} p q^\top W_{j-1\sim x}$. Observe that $p \in a_{lji} \subseteq A_{lji} \subseteq \operatorname{null} W_{l+1\sim j}$ and $q \in b_{k,j-1,h} \subseteq B_{k,j-1,h} \subseteq \operatorname{null} W^\top_{j-1\sim h-1}$. Therefore, if $y > l$ then $W_{y\sim j} p = \mathbf{0}$; symmetrically, if $x < h$ then $W^\top_{j-1\sim x} q = \mathbf{0}$. Thus if $\epsilon = 0$ or $y > l$ or $x < h$, then $\Delta W_{y\sim x} = 0$ and $W'_{y\sim x} = W_{y\sim x}$.

By Lemma 6, $p \notin \operatorname{null} W_{l\sim j}$ and $q \notin \operatorname{null} W^\top_{j-1\sim h}$. Hence if $\epsilon \neq 0$ and $j \in [x+1, y]$ and $y \leq l$ and $x \geq h$, then $W_{y\sim j} p \neq \mathbf{0}$, $W^\top_{j-1\sim x} q \neq \mathbf{0}$, $\Delta W_{y\sim x} \neq 0$, and thus $W'_{y\sim x} \neq W_{y\sim x}$. □

The next lemma is preparation for the lemma that follows it.

Lemma 15. *For every vector $p \in a_{lji} \setminus \{\mathbf{0}\}$, $p \notin \operatorname{null} W_{l\sim j} + \operatorname{col} W_{j\sim i-1}$. Similarly, for every vector $q \in b_{k,j-1,h} \setminus \{\mathbf{0}\}$, $q \notin \operatorname{row} W_{k+1\sim j-1} + \operatorname{null} W^\top_{j-1\sim h}$.*

Proof. Observe that $\operatorname{null} W_{l\sim j} = A_{l-1,j,j}$ and $\operatorname{col} W_{j\sim i-1} = A_{L,j,i-1}$. By Lemma 4, $\mathcal{A}_{Ljj}$ is a prebasis for $A_{Ljj} = \mathbb{R}^{d_j}$, $\mathcal{A}_{l-1,j,j} \subseteq \mathcal{A}_{Ljj}$ is a prebasis for $A_{l-1,j,j} \subseteq \mathbb{R}^{d_j}$, and $\mathcal{A}_{L,j,i-1} \subseteq \mathcal{A}_{Ljj}$ is a prebasis for $A_{L,j,i-1} \subseteq \mathbb{R}^{d_j}$. As $a_{lji} \in \mathcal{A}_{Ljj}$ and $\mathcal{A}_{l-1,j,j} \cup \mathcal{A}_{L,j,i-1} \subseteq \mathcal{A}_{Ljj}$ but $a_{lji} \notin \mathcal{A}_{l-1,j,j} \cup \mathcal{A}_{L,j,i-1}$, the fact that $\mathcal{A}_{Ljj}$ is a prebasis implies that a_{lji} is linearly independent of the subspaces in $\mathcal{A}_{l-1,j,j} \cup \mathcal{A}_{L,j,i-1}$, so $a_{lji} \cap (A_{l-1,j,j} + A_{L,j,i-1}) = \{\mathbf{0}\}$. Hence, for every $p \in a_{lji} \setminus \{\mathbf{0}\}$, $p \notin A_{l-1,j,j} + A_{L,j,i-1} = \operatorname{null} W_{l\sim j} + \operatorname{col} W_{j\sim i-1}$.

A symmetric argument shows the second claim. □

The next lemma identifies which subsequence matrices do or do not have new vectors appear in their rowspaces or columnspaces.

Lemma 16. *Given $L \geq y \geq x \geq 0$, $\operatorname{col} \Delta W_{y\sim x} \subseteq \operatorname{col} W_{y\sim x}$ (equivalently, $\operatorname{col} W'_{y\sim x} \subseteq \operatorname{col} W_{y\sim x}$) if and only if $\epsilon = 0$ or $j \notin [x+1, y]$ or $y > l$ or $x < h$ or $x \geq i$. Symmetrically, $\operatorname{row} \Delta W_{y\sim x} \subseteq \operatorname{row} W_{y\sim x}$ (equivalently, $\operatorname{row} W'_{y\sim x} \subseteq \operatorname{row} W_{y\sim x}$) if and only if $\epsilon = 0$ or $j \notin [x+1, y]$ or $y > l$ or $x < h$ or $y \leq k$.*

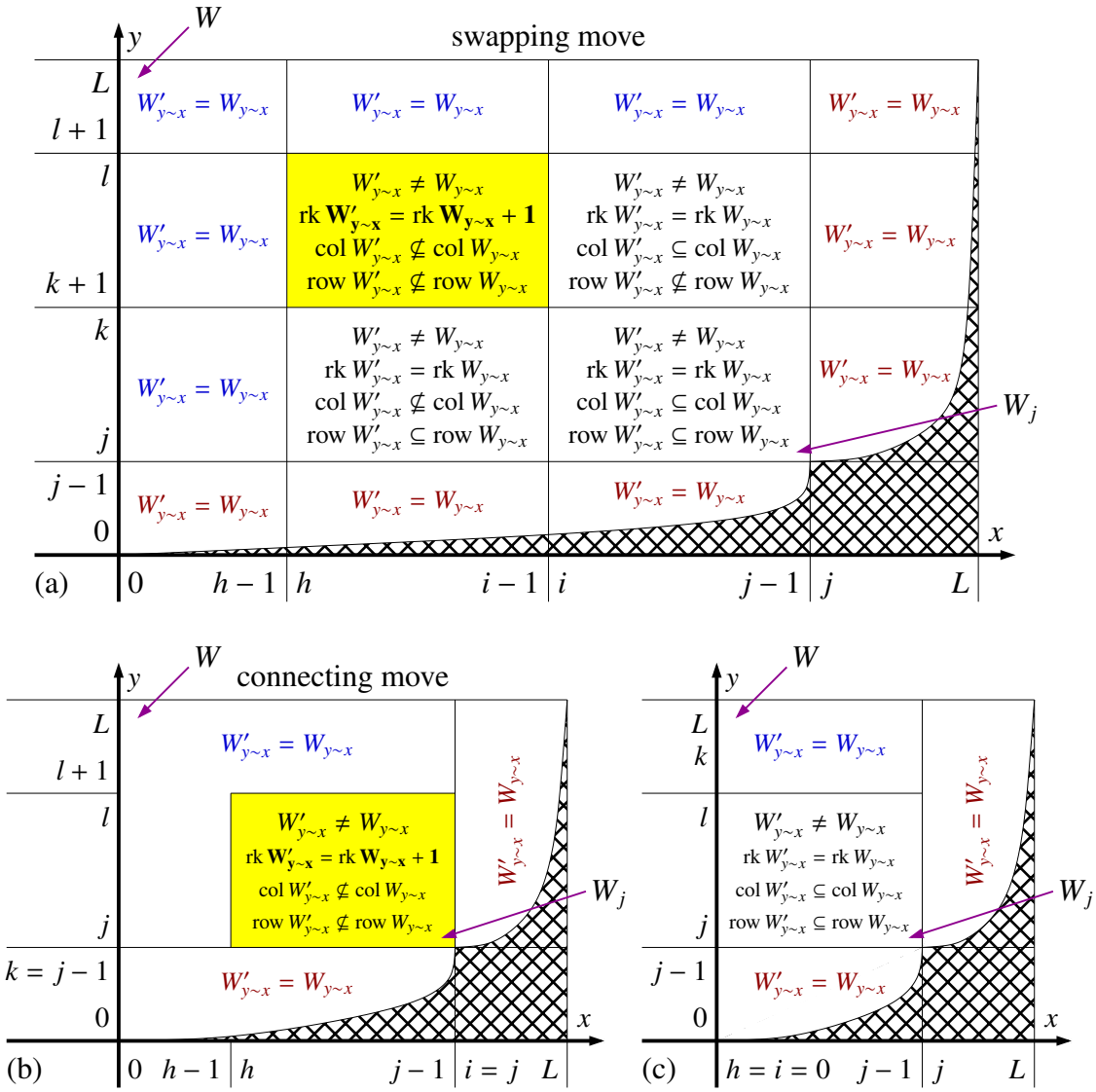


Table 4: The influence of an infinitesimal one-matrix move (i.e., ϵ is sufficiently small) in which the factor matrix W_j undergoes a displacement $\Delta W_j \in o_{lkjih}$. The effects on the subsequence matrix $W_{y\sim x}$ are listed for every y and x with $y \geq x$. These tables are triangular, though it's not obvious at first: the hatched region represents an unused zone where $y < x$. A yellow rectangle indicates which subsequence matrices increase in rank, constituting a combinatorial (connecting or swapping) move. The black font indicates where $W'_{y\sim x} \neq W_{y\sim x}$. The red font indicates where $W'_{y\sim x} = W_{y\sim x}$ because the matrix W_j is not a factor in $W_{y\sim x}$. The blue font indicates where $W'_{y\sim x} = W_{y\sim x}$ for deeper reasons. (a) Table for the case where $L > l > k > j-1$ and $j > i > h > 0$. An example of a swapping move. (b) The third row disappears if $k = j-1$, and the third column disappears if $i = j$. When both identities hold, the move is a connecting move. (c) The second row disappears if $k \geq l$, and the second column disappears if $i \leq h$. If either inequality holds, the move is not a combinatorial move. The first column disappears if $h = 0$. (The first row disappears if $l = L$, not shown. If $h = 0$ and $l = L$, then $W' \neq W$ and we move off the fiber.)

Proof. If $\epsilon = 0$ or $j \notin [x+1, y]$ or $y > l$ or $x < h$, then $\Delta W_{y\sim x} = 0$ by Lemma 14 and the result follows. Henceforth, assume that $\epsilon \neq 0$ and $j \in [x+1, y]$ and $y \leq l$ and $x \geq h$.

As $\Delta W_{y\sim x} = \epsilon W_{y\sim j} p q^\top W_{j-1\sim x}$, $\operatorname{col} \Delta W_{y\sim x}$ is either $\{\mathbf{0}\}$ or the line spanned by the vector $W_{y\sim j} p$. As $p \in a_{lji} \subseteq \operatorname{col} W_{j\sim i}$, $W_{y\sim j} p \in \operatorname{col} W_{y\sim i}$. If $x \geq i$ then $W_{y\sim j} p \in \operatorname{col} W_{y\sim x}$, so $\operatorname{col} \Delta W_{y\sim x} \subseteq \operatorname{col} W_{y\sim x}$ as claimed.

Symmetrically, $\operatorname{row} \Delta W_{y\sim x}$ is either $\{\mathbf{0}\}$ or the line spanned by the vector $W_{j-1\sim x}^\top q$. As $q \in b_{k,j-1,h} \subseteq \operatorname{row} W_{k\sim j-1}$, $W_{j-1\sim x}^\top q \in \operatorname{row} W_{k\sim x}$. If $y \leq k$ then $W_{j-1\sim x}^\top q \in \operatorname{row} W_{y\sim x}$, so $\operatorname{row} \Delta W_{y\sim x} \subseteq \operatorname{row} W_{y\sim x}$ as claimed.

By Lemma 15, $p \notin \operatorname{null} W_{l\sim j} + \operatorname{col} W_{j\sim i-1}$ and $q \notin \operatorname{row} W_{k+1\sim j-1} + \operatorname{null} W_{j-1\sim h}^\top$. Therefore, $W_{l\sim j} p \notin \operatorname{col} W_{l\sim i-1}$ and $W_{j-1\sim h}^\top q \notin \operatorname{row} W_{k+1\sim h}$. As $y \leq l$, $W_{y\sim j} p \notin \operatorname{col} W_{y\sim i-1}$. As $x \geq h$, $W_{j-1\sim x}^\top q \notin \operatorname{row} W_{k+1\sim x}$.

If $x < i$, $\operatorname{col} W_{y\sim i-1} \supseteq \operatorname{col} W_{y\sim x}$ and thus $W_{y\sim j} p \notin \operatorname{col} W_{y\sim x}$. We have $W_{j-1\sim x}^\top q \neq \mathbf{0}$ because $q \notin \operatorname{null} W_{j-1\sim h}^\top$, which means that $W_{x\sim h}^\top W_{j-1\sim x}^\top q \neq \mathbf{0}$. Hence $\operatorname{col} \Delta W_{y\sim x}$ is the line spanned by the vector $W_{y\sim j} p$, which is not in $\operatorname{col} W_{y\sim x}$, so $\operatorname{col} \Delta W_{y\sim x} \nsubseteq \operatorname{col} W_{y\sim x}$ as claimed.

If $y > k$, $\operatorname{row} W_{k+1\sim x} \supseteq \operatorname{row} W_{y\sim x}$ and thus $W_{j-1\sim x}^\top q \notin \operatorname{row} W_{y\sim x}$. We have $W_{y\sim j} p \neq \mathbf{0}$ because $p \notin \operatorname{null} W_{l\sim j}$, which means that $W_{l\sim y} W_{y\sim j} p \neq \mathbf{0}$. Hence $\operatorname{row} \Delta W_{y\sim x}$ is the line spanned by the vector $W_{j-1\sim x}^\top q$, which is not in $\operatorname{row} W_{y\sim x}$, so $\operatorname{row} \Delta W_{y\sim x} \nsubseteq \operatorname{row} W_{y\sim x}$ as claimed. □

The next lemma addresses the crucial question of which moves can change the rank of a subsequence matrix—that is, which moves are combinatorial. It begins with a general statement for both finite and infinitesimal moves, then gives a stronger statement for infinitesimal moves. This is the only result in this section (Section 6.2) that treats infinitesimal moves differently than finite ones.

Lemma 17. *Given $L \geq y \geq x \geq 0$, for all $\epsilon \in \mathbb{R}$, $\operatorname{rk} W'_{y\sim x} \leq \operatorname{rk} W_{y\sim x} + 1$. Moreover, if $y > l$ or $y \leq k$ or $x \geq i$ or $x < h$, then $\operatorname{rk} W'_{y\sim x} \leq \operatorname{rk} W_{y\sim x}$.*

Moreover, there exists an $\hat{\epsilon} > 0$ such that for all $\epsilon \in (-\hat{\epsilon}, \hat{\epsilon})$, $\operatorname{rk} W'_{y\sim x} = \operatorname{rk} W_{y\sim x} + 1$ if $\epsilon \neq 0$ and $l \geq y > k$ and $i > x \geq h$, and $\operatorname{rk} W'_{y\sim x} = \operatorname{rk} W_{y\sim x}$ otherwise.

Proof. The displacement $\Delta W_{y\sim x} = (\epsilon W_{y\sim j} p)(q^\top W_{j-1\sim x})$ is an outer product of two vectors, so its rank is one or zero and $\operatorname{rk} W'_{y\sim x} \leq \operatorname{rk} W_{y\sim x} + 1$. If $y > l$ or $y \leq k$, then $\operatorname{row} W'_{y\sim x} \subseteq \operatorname{row} W_{y\sim x}$ by Lemma 16, so $\operatorname{rk} W'_{y\sim x} \leq \operatorname{rk} W_{y\sim x}$. If $x \geq i$ or $x < h$, then $\operatorname{col} W'_{y\sim x} \subseteq \operatorname{col} W_{y\sim x}$ by Lemma 16, and again $\operatorname{rk} W'_{y\sim x} \leq \operatorname{rk} W_{y\sim x}$. If any of those conditions hold ($y > l$ or $y \leq k$ or $x \geq i$ or $x < h$) and moreover ϵ is sufficiently small, then $\operatorname{rk} W'_{y\sim x} = \operatorname{rk} W_{y\sim x}$, as decreasing the rank requires some finite displacement. If $\epsilon = 0$, then $W'_{y\sim x} = W_{y\sim x}$.

If $\epsilon \neq 0$ and $l \geq y > k$ and $i > x \geq h$, then we have $j \in [x+1, y]$ because $j \geq i \geq x+1$ and $j \leq k+1 \leq y$. Then by Lemma 16, $\operatorname{col} \Delta W_{y\sim x} \nsubseteq \operatorname{col} W_{y\sim x}$ and $\operatorname{row} \Delta W_{y\sim x} \nsubseteq \operatorname{row} W_{y\sim x}$. Therefore, if ϵ is sufficiently small, then $\operatorname{rk} W'_{y\sim x} = \operatorname{rk} W_{y\sim x} + 1$. □

6.3 Infinitesimal Combinatorial Moves

Recall that a move is *combinatorial* if it changes the rank of one or more of the subsequence matrices. An infinitesimal move cannot decrease any matrix rank, but it might increase one or more ranks. Hence, the *infinitesimal combinatorial moves* are the infinitesimal moves that increase some subsequence matrix rank. It follows from Lemma 17 that these are the moves in which a displacement ΔW_j is chosen from a subspace o_{lkjih} such that $l > k$ and $i > h$. Lemma 17 shows that the subsequence matrices whose ranks increase are $W_{y\sim x}$ for all $y \in [k+1, l]$ and $x \in [h, i-1]$ (as Table 4 illustrates). Their ranks all increase by the same amount: the rank of ΔW_j.

Interestingly, a single move may change the ranks of many subsequence matrices, but at most four interval multiplicities change. Recall the identity (8), $\omega_{ts} = \operatorname{rk} W_{t\sim s} - \operatorname{rk} W_{t\sim s-1} - \operatorname{rk} W_{t+1\sim s} + \operatorname{rk} W_{t+1\sim s-1}$. If all four ranks increase by $\operatorname{rk} \Delta W_j$, or exactly two ranks with opposite signs do, then ω_{ts} does not change.

It is straightforward to check that ω_{kh} and ω_{li} decrease by $\operatorname{rk} \Delta W_j$, ω_{lh} and ω_{ki} increase by $\operatorname{rk} \Delta W_j$, and no other interval multiplicity changes. Hence, the integer multiplicities encode the changes produced by an infinitesimal combinatorial move more elegantly than the rank list does.

By the definition of Θ_{O}, $k + 1 \geq j \geq i$. In the special case where $k + 1 = j = i$, we call the infinitesimal combinatorial move a *connecting move*. In a connecting move, ω_{ki} does not exist (as $k < i$) and only three interval multiplicities change. Figure 8 illustrates two examples of connecting moves and offers an intuitive way to interpret them: a connecting move deletes $\operatorname{rk} \Delta W_j$ copies of the interval $[h, k] = [h, j-1]$ and $\operatorname{rk} \Delta W_j$ copies of the interval $[i, l] = [j, l]$, and replaces them with $\operatorname{rk} \Delta W_j$ copies of the interval $[h, l]$. We think of this as connecting the intervals $[h, j-1]$ and $[j, l]$ together with an added edge $[j-1, j]$ to create an interval $[h, l]$; hence the name "connecting move." (There is much intuition that can be gleaned from a careful study of the figure that is hard to explain in words.)

A *swapping move* is an infinitesimal combinatorial move with $k \geq i$, which changes four interval multiplicities. Figure 9 illustrates two examples of swapping moves. A swapping move splices $\operatorname{rk} \Delta W_j$ copies of the interval $[h, k]$ with $\operatorname{rk} \Delta W_j$ copies of the interval $[i, l]$, thereby replacing them with $\operatorname{rk} \Delta W_j$ copies of the interval $[h, l]$ (which is longer than both of the replaced intervals) and $\operatorname{rk} \Delta W_j$ copies of the interval $[i, k]$ (which is shorter than both).

The ideas of connecting and swapping moves, along with Figures 8 and 9, expose much intuition about how strata are connected to each other. An infinitesimal move reflects the ways that an infinitesimal perturbation of a point in weight space can move you from one stratum to another stratum; the former stratum is a subset of the closure of the latter stratum. One could argue that the strata would better be indexed by the interval multiplicities than the rank lists, because the interval multiplicities make it easier to see how you can move from one stratum to another.

We define the following sets of one-matrix subspaces that correspond to combinatorial moves, connecting moves, and swapping moves.

$$\begin{aligned}
\Theta_{\mathrm{O}}^{\mathrm{comb}} &= \{\phi_{lkjih} \in \Theta_{\mathrm{O}} : l > k \text{ and } i > h\} = \{\phi_{lkjih} \neq \{\mathbf{0}\} : L \geq l \geq k+1 \geq j \geq i > h \geq 0\}, \\
\Theta_{\mathrm{O}}^{\mathrm{conn}} &= \{\phi_{lkjih} \in \Theta_{\mathrm{O}} : l \geq k+1 = j = i > h\} = \{\phi_{l,j-1,j,j,h} \neq \{\mathbf{0}\} : L \geq l \geq j > h \geq 0\}, \\
\Theta_{\mathrm{O}}^{\mathrm{swap}} &= \{\phi_{lkjih} \in \Theta_{\mathrm{O}} : l > k \geq i > h\} = \{\phi_{lkjih} \neq \{\mathbf{0}\} : L \geq l > k \geq i > h \geq 0 \text{ and } k+1 \geq j \geq i\}.
\end{aligned}$$

6.4 The Differential Map $\mathrm{d}\mu$ and its Nullspace

Imagine you are standing at a point θ on a fiber $\mu^{-1}(W)$. As μ is a polynomial function of θ, μ is smooth, but the fiber might not be locally manifold at θ, in which case the fiber is not smooth at θ. Nevertheless, if you walk on a path on the fiber starting from θ, your initial direction of motion $\Delta\theta$ is necessarily one along which the directional derivative $\mu'_{\Delta\theta}(\theta)$ is zero. (Note that this directional derivative is a matrix; all of its components must be zero.) But the converse does not hold—not every direction with derivative zero necessarily is associated with some path on the fiber! For example, if you are standing at the origin ($\theta \in S_{00}$) in Figure 2, the directional derivative of μ is zero for *every* direction in weight space, but only some directions stay on the fiber.

To better understand these derivatives, researchers use concepts from differential geometry. Given a neural network architecture $\mu : \mathbb{R}^{d_\theta} \to \mathbb{R}^{d_h \times d_0}$ and a specific weight vector $\theta \in \mathbb{R}^{d_\theta}$, the *differential map* $\mathrm{d}\mu(\theta)$:

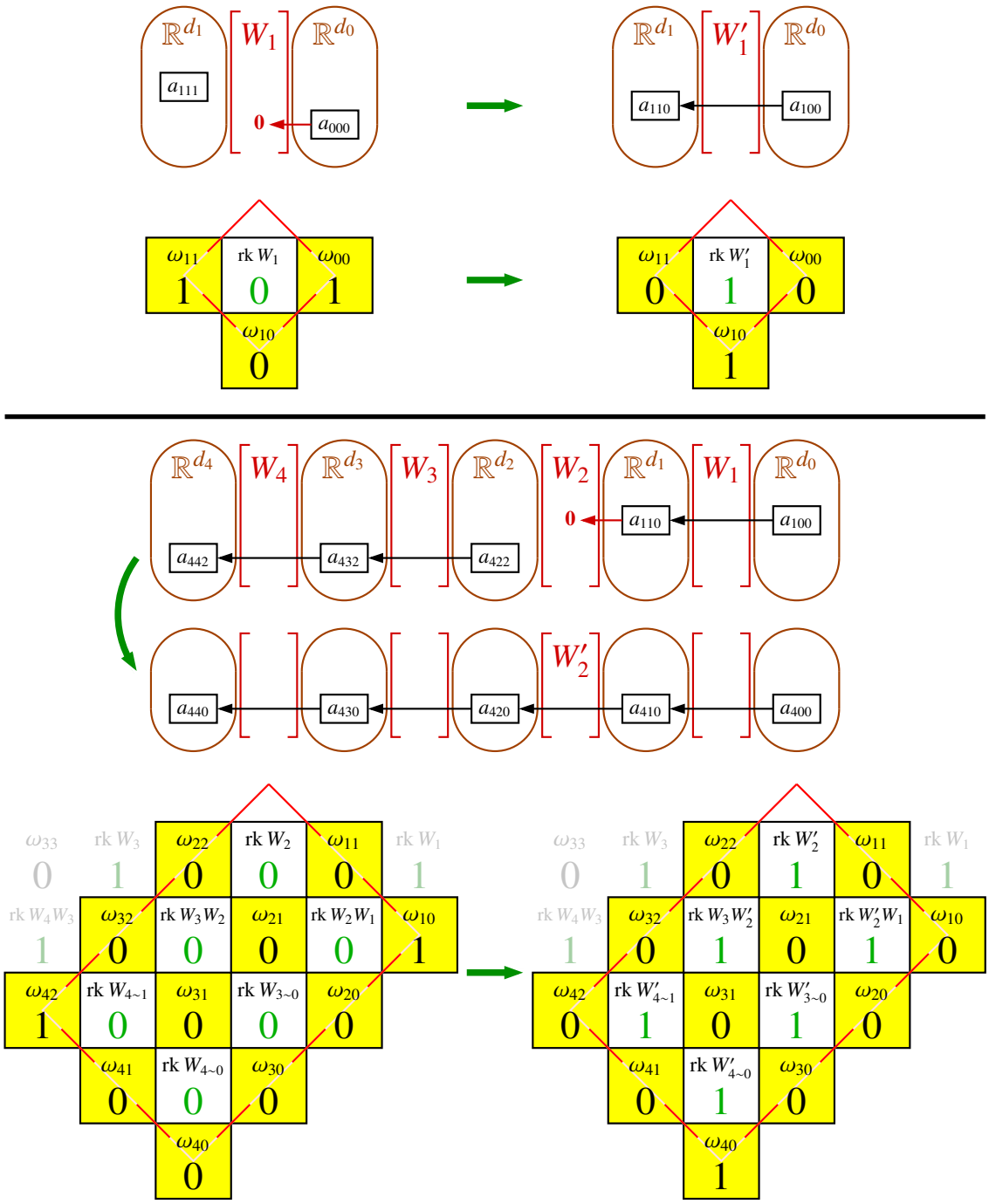


Figure 8: Two examples of connecting moves. The top example is the simplest possible example: W_1 has been perturbed to increase its rank by one. In the bottom example, W_2 has been perturbed. In both examples, the perturbation of W_j causes two intervals $[h, j-1]$ and $[j, l]$ to be replaced by a single interval $[h, l]$. Three interval multiplicities change, at three of the four corners of the red rectangle: ω_{lj} and $\omega_{j-1,h}$ decrease by one, and ω_{lh} increases by one. The ranks of the subsequence matrices $W_{y\sim x}$ increase by one for all $y \in [j, l]$ and $x \in [h, j-1]$ (the ranks inside the red rectangle, including rk W_j). *Outside* the red rectangle, all interval multiplicities and matrix ranks are unchanged.

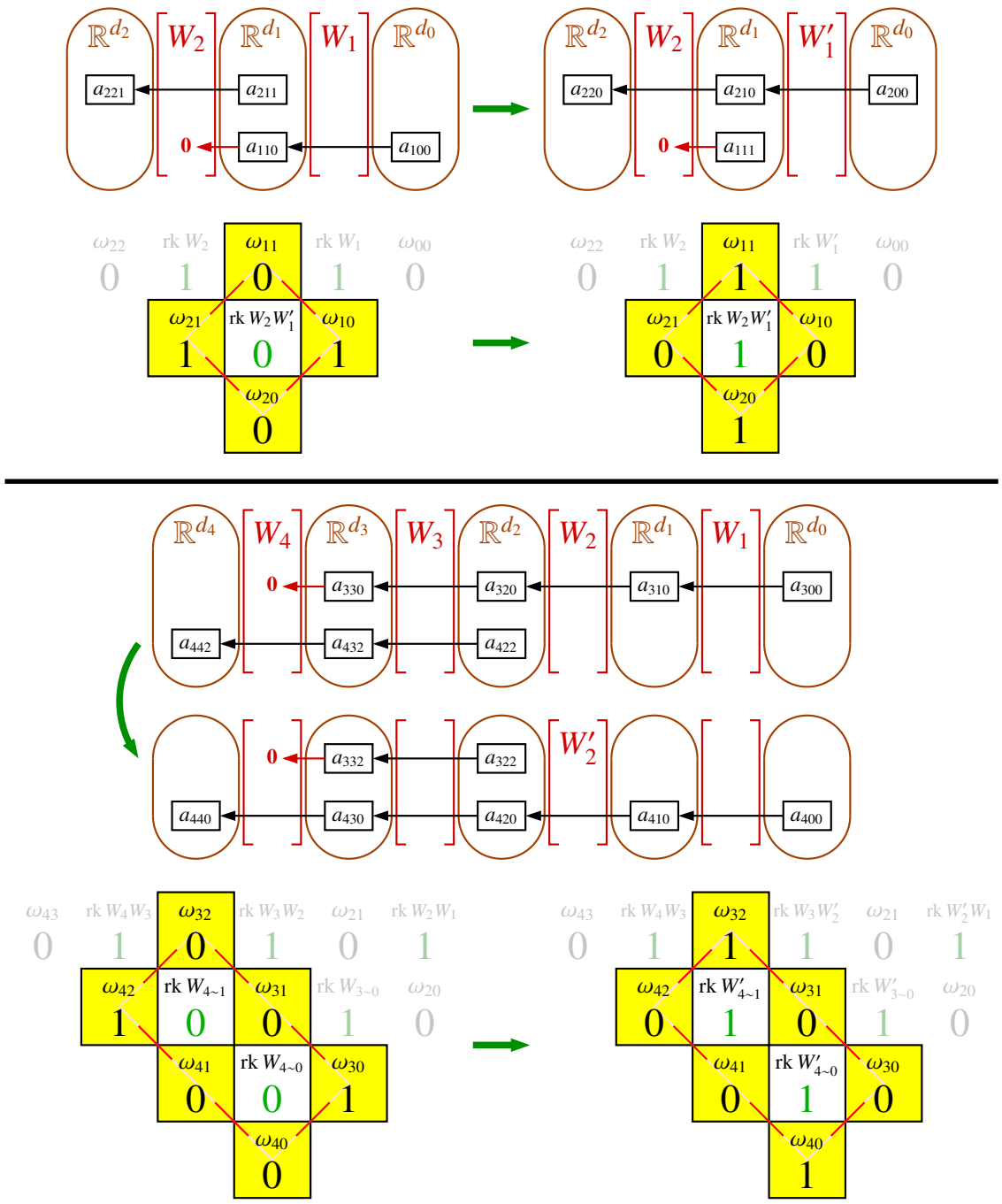


Figure 9: Two examples of swapping moves. In the top example—the simplest possible example—either W_1 or W_2 may be perturbed to cause the move. In the bottom example, any of of W_2, W_3, or W_4 may have been perturbed. Two intervals $[h, k]$ and $[i, l]$ are replaced by an interval $[h, l]$, longer than both original intervals, and an interval $[i, k]$, shorter than both. Four interval multiplicities change, at the four corners of the red rectangle: ω_{kh} and ω_{li} decrease by one, and ω_{lh} and ω_{ki} increase by one. The ranks of the subsequence matrices $W_{y\sim x}$ increase by one for all $y \in [k+1, l]$ and $x \in [h, i-1]$ (the ranks inside the red rectangle). *Outside* the red rectangle, all interval multiplicities and matrix ranks are unchanged.

$\mathbb{R}^{d_\theta} \to \mathbb{R}^{d_h \times d_0}$ is a linear map from weight space to the space of the matrix W. We emphasize the linearity; think of the differential map as the linear term in a Taylor expansion of μ about θ. In general, we will write its argument as $\Delta\theta$, and apply the map as $\Delta W = \mathrm{d}\mu(\theta)(\Delta\theta)$. The notations ΔW and $\Delta\theta$ reflect a natural interpretation in terms of perturbations: if you are at a point θ in weight space, yielding a matrix $W = \mu(\theta)$, then you perturb θ by an infinitesimal displacement $\Delta\theta$, the matrix W is perturbed by an infinitesimal ΔW.

The bare form $\mathrm{d}\mu$ denotes a map from a weight vector θ to a linear map. This might seem confusing if you haven't seen it before—a map that produces a map—and it accounts for the odd notation $\mathrm{d}\mu(\theta)(\Delta\theta)$.

Let $\Delta\theta = (\Delta W_L, \Delta W_{L-1}, \ldots, \Delta W_1) \in \mathbb{R}^{d_\theta}$ be a weight perturbation. By the chain rule, the value of the differential map for μ at a fixed weight vector θ is

$$\mathrm{d}\mu(\theta)(\Delta\theta) = \sum_{j=1}^{L} W_{L\sim j} \Delta W_j W_{j-1\sim 0} = \Delta W_L W_{L-1\sim 0} + W_L \Delta W_{L-1} W_{L-2\sim 0} + \ldots + W_{L\sim 1} \Delta W_1. \tag{9}$$

If you walk from θ along a smooth path on the fiber $\mu^{-1}(W)$, your initial direction of motion is in the nullspace of $\mathrm{d}\mu(\theta)$, defined to be

$$\operatorname{null} \mathrm{d}\mu(\theta) = \{\Delta\theta \in \mathbb{R}^{d_\theta} : \mathrm{d}\mu(\theta)(\Delta\theta) = 0\}.$$

For that reason, we want to specify a prebasis for $\operatorname{null} \mathrm{d}\mu(\theta)$, which we call the *fiber prebasis*. In Section 6.1 we characterized all the one-matrix moves that stay on the fiber. Recall that we define a one-matrix move to be a move with displacement $\Delta\theta \in \phi_{lkjih}$ for some subspace $\phi_{lkjih} \in \Theta_{\mathrm{O}}$, and we define $\Theta_{\mathrm{O}}^{\mathrm{fiber}} \subseteq \Theta_{\mathrm{O}}$ to contain all the subspaces representing moves that stay on the fiber.

Lemma 18. *For every subspace $\phi_{lkjih} \in \Theta_{\mathrm{O}}^{\mathrm{fiber}}$, $\phi_{lkjih} \subseteq \operatorname{null} \mathrm{d}\mu(\theta)$.*

Proof. Consider a perturbation $\Delta\theta \in \phi_{lkjih}$; we can write $\Delta\theta = (\ldots, 0, \Delta W_j, 0, \ldots)$. By the definition of $\Theta_{\mathrm{O}}^{\mathrm{fiber}}$, either $L > l$ or $h > 0$. In the former case, $W_{L\sim j}\Delta W_j = 0$ because $\operatorname{row} \Delta W_j \subseteq a_{lji} \subseteq A_{lji} \subseteq \operatorname{null} W_{l+1\sim j}$. In the latter case, $\Delta W_j W_{j-1\sim 0} = 0$ because $\operatorname{col} \Delta W_j \subseteq b_{k,j-1,h} \subseteq B_{k,j-1,h} \subseteq \operatorname{null} W_{j-1\sim h-1}^{\top}$. In both cases, by the formula (9), $\mathrm{d}\mu(\theta)(\Delta\theta) = W_{L\sim j}\Delta W_j W_{j-1\sim 0} = 0$. Hence $\phi_{lkjih} \subseteq \operatorname{null} \mathrm{d}\mu(\theta)$. □

Usually $\Theta_{\mathrm{O}}^{\mathrm{fiber}}$ does not suffice to span $\operatorname{null} \mathrm{d}\mu(\theta)$. To give a complete prebasis for $\operatorname{null} \mathrm{d}\mu(\theta)$, usually we must add some two-matrix moves. To give a prebasis for the stratum containing the point θ, usually we must remove some more one-matrix moves (those that move off the stratum) and add some more two-matrix moves.

Recall that a two-matrix move uses two nonzero displacement matrices ΔW_{j+1} and ΔW_j. In a one-matrix move that stays on the fiber, every term in the summation (9) is zero. In a two-matrix move (in its differential form), exactly two terms in the summation (9) are nonzero, offsetting each other so the sum is zero.

6.5 Two-Matrix Moves

A *finite two-matrix move* is a move that selects an invertible $d_j \times d_j$ matrix M and modifies the factor matrices W_{j+1} and W_j by setting $W'_{j+1} = W_{j+1}M$ and $W'_j = M^{-1}W_j$. The other factor matrices do not change: $W'_z = W_z$ for all $z \notin \{j, j+1\}$. Clearly, all finite two-matrix moves stay on the fiber: $W' = \mu(\theta') = \mu(\theta) = W$. Moreover, a finite two-matrix move does not change the rank of any subsequence matrix, so the move stays on the same stratum. One way to think of this move: an invertible linear transformation changes how hidden layer j represents information, without otherwise changing anything that the network computes.

We also define a differential (infinitesimal) version of two-matrix moves. Given a finite two-matrix move from θ to θ' as described above, there is typically no straight path on the fiber from θ to θ', but there is a natural curved path. Intuitively, we want a differential move to capture the initial direction of a path on the fiber, without caring where the path ends. That initial direction of movement lies in the nullspace of the differential map $\mathrm{d}\mu(\theta)$. Therefore, we define a differential move to have a displacement $\Delta\theta \in \operatorname{null} \mathrm{d}\mu(\theta)$. Because of the fiber's curvature, the point $\theta + \Delta\theta$ is *not* necessarily on the fiber—that is, often $\mu(\theta + \Delta\theta)$ does not equal W.

Consider a finite two-matrix move where $M = I + \epsilon K$ for an arbitrary, nonzero $d_j \times d_j$ matrix K. For a sufficiently small ϵ, M is invertible, so we can draw a smooth path on the fiber leaving θ by varying ϵ from zero to some small value. (The curved grid lines in Figure 1 are examples of such paths.) The entire path lies on the same stratum as θ. To find the line tangent to this path at θ and to generate a differential two-matrix move, observe that for a small ϵ, $(I + \epsilon K)^{-1} = I - \epsilon K + \epsilon^2 K^2 - \epsilon^3 K^3 + \ldots$, so

$$\begin{aligned}
\frac{\mathrm{d}}{\mathrm{d}\epsilon} W'_{j+1} &= \frac{\mathrm{d}}{\mathrm{d}\epsilon}\Big(W_{j+1}(I + \epsilon K)\Big) = W_{j+1}K \quad \text{and} \\
\frac{\mathrm{d}}{\mathrm{d}\epsilon} W'_j\Big|_{\epsilon=0} &= \frac{\mathrm{d}}{\mathrm{d}\epsilon}\Big((I + \epsilon K)^{-1} W_j\Big)\Big|_{\epsilon=0} = \frac{\mathrm{d}}{\mathrm{d}\epsilon}\Big((I - \epsilon K + \epsilon^2 K^2 - \epsilon^3 K^3 + \ldots) W_j\Big)\Big|_{\epsilon=0} = -KW_j.
\end{aligned}$$

Therefore, we are interested in differential moves with a displacement

$$\Delta\theta \propto (0, 0, \ldots, 0, W_{j+1}K, -KW_j, 0, \ldots, 0)$$

where the components are positioned so that $\Delta W_{j+1} \propto W_{j+1}K$ and $\Delta W_j \propto -KW_j$. As $\Delta\theta$ is tangent to the path at θ (assuming $\Delta\theta \neq \mathbf{0}$), it is tangent to the stratum at θ.

A useful way to define subspaces of differential two-matrix moves is to consider matrices $K \in a_{lji} \otimes b_{kjh}$. Recall that this means that $\operatorname{col} K \subseteq a_{lji}$ and $\operatorname{row} K \subseteq b_{kjh}$. For all l, k, j, i, h that satisfy $L > j > 0$, $L \geq l \geq j \geq i \geq 0$, and $L \geq k \geq j \geq h \geq 0$, the *two-matrix subspaces* are

$$\tau_{lkjih} = \{(0, 0, \ldots, 0, W_{j+1}K, -KW_j, 0, \ldots, 0) : K \in a_{lji} \otimes b_{kjh}\}. \tag{10}$$

The dimension of τ_{lkjih} is $(\dim a_{lji}) \cdot (\dim b_{kjh}) = \omega_{li}\,\omega_{kh}$. We define a *differential two-matrix move* to be a move with a displacement $\Delta\theta \in \tau_{lkjih}$. All such displacements lie in the nullspace of the differential map.

Lemma 19. *For all l, k, j, i, h that satisfy $L > j > 0$, $L \geq l \geq j \geq i \geq 0$, and $L \geq k \geq j \geq h \geq 0$, $\tau_{lkjih} \subseteq \operatorname{null} \mathrm{d}\mu(\theta)$.*

Proof. Consider a perturbation $\Delta\theta \in \tau_{lkjih}$. By the definition (10), there exists some $K \in a_{lji} \otimes b_{kjh}$ such that $\Delta W_{j+1} = W_{j+1}K$ and $\Delta W_j = -KW_j$; the other matrices in $\Delta\theta$ are zeros. By the formula (9), $\mathrm{d}\mu(\theta)(\Delta\theta) = W_{L\sim j+1}\Delta W_{j+1}W_{j\sim 0} + W_{L\sim j}\Delta W_j W_{j-1\sim 0} = W_{L\sim j+1}W_{j+1}KW_{j\sim 0} - W_{L\sim j}KW_jW_{j-1\sim 0} = 0$. Hence $\Delta\theta \in \operatorname{null} \mathrm{d}\mu(\theta)$ and $\tau_{lkjih} \subseteq \operatorname{null} \mathrm{d}\mu(\theta)$. □

Let S be the stratum of $\mu^{-1}(W)$ that contains θ, where $W = \mu(\theta)$. Let $T_\theta S$ be the tangent subspace of S at θ. Every differential two-matrix move has a displacement $\Delta\theta$ that lies in $T_\theta S$—that is, $\Delta\theta$ is tangent to S at θ.

Lemma 20. *For all l, k, j, i, h that satisfy $L > j > 0$, $L \geq l \geq j \geq i \geq 0$, and $L \geq k \geq j \geq h \geq 0$, $\tau_{lkjih} \subseteq T_\theta S$.*

Proof. Consider a perturbation $\Delta\theta \in \tau_{lkjih}$. There exists some $K \in a_{lji} \otimes b_{kjh}$ such that $\Delta W_{j+1} = W_{j+1}K$ and $\Delta W_j = -KW_j$. Consider the path

$$P = \{(W_L, \ldots, W_{j+2}, W_{j+1}(I + \epsilon K), (I + \epsilon K)^{-1}W_j, W_{j-1}, \ldots, W_1 : \epsilon \in [0, \epsilon']\}$$

where $\epsilon' > 0$ is sufficiently small that $I + \epsilon K$ is invertible for all $\epsilon \in [0, \epsilon']$. The path P is connected and smooth, and θ is one of its endpoints. It satisfies $P \subset S$, as all points $\theta' \in P$ satisfy $\mu(\theta') = \mu(\theta)$ and have the same subsequence matrix ranks as θ. As we have seen, $\mathrm{d}\theta/\mathrm{d}\epsilon|_{\epsilon=0} = \Delta\theta$, so $\Delta\theta$ is tangent to the smooth path P at θ, which implies that $\Delta\theta$ is tangent to S at θ. Hence $\Delta\theta \in T_\theta S$ and $\tau_{lkjih} \subseteq T_\theta S$. □

Some of the two-matrix subspaces are redundant with the one-matrix subspaces. If $l = j$ then $W_{j+1}K = 0$ (because $\operatorname{col} K \subseteq a_{jji} \subseteq A_{jji} \subseteq \operatorname{null} W_{j+1}$), so τ_{jkjih} degenerates to a one-matrix subspace. Likewise, if $j = h$ then $KW_j = 0$ (because $\operatorname{row} K \subseteq b_{kjj} \subseteq B_{kjj} \subseteq \operatorname{null} W_j^\top$), so τ_{lkjij} degenerates to a one-matrix subspace. Degenerate two-matrix subspaces represent directions that are already spanned by $\Theta_{\mathrm{O}}^{\mathrm{fiber}}$, so from here on we assume that $l > j > h$. It is useful to define the set of two-matrix subspaces that are not one-matrix subspaces,

$$\Theta_{\mathrm{T}} = \{\tau_{lkjih} \neq \{\mathbf{0}\} : L \geq l > j \geq i \geq 0 \text{ and } L \geq k \geq j > h \geq 0\}.$$

The following lemma shows that the two matrices in a two-matrix subspace in Θ_{T} cannot vary independently.

Lemma 21. *Consider a subspace $\tau_{lkjih} \in \Theta_{\mathrm{T}}$ and a displacement $\Delta\theta \in \tau_{lkjih}$; thus we can write $\Delta\theta = (\ldots, 0, \Delta W_{j+1}, \Delta W_j, 0, \ldots)$. Then ΔW_{j+1} and ΔW_j are either both zero or both nonzero.*

Proof. From the definition (10) of τ_{lkjih}, there is a $K \in a_{lji} \otimes b_{kjh}$ such that $\Delta W_{j+1} = W_{j+1}K$ and $\Delta W_j = -KW_j$. If $K = 0$ then $\Delta W_{j+1} = 0$ and $\Delta W_j = 0$. By the definition of Θ_{T}, $l > j > h$. So if $K \neq 0$ then $\operatorname{col} K \not\subseteq \operatorname{null} W_{j+1}$ and $\operatorname{row} K \not\subseteq \operatorname{null} W_j^\top$; hence $\Delta W_{j+1} \neq 0$ and $\Delta W_j \neq 0$. □

For our purposes, it is best if we choose a_{lji} and b_{kjh} to be flow prebasis subspaces (rather than choosing, say, the standard prebases), as Lemma 7 says we always can, so that $\operatorname{col} W_{j+1}K \subseteq a_{l,j+1,i}$ and $\operatorname{row} KW_j \subseteq b_{k,j-1,h}$. Then a displacement $\Delta\theta \in \tau_{lkjih}$ satisfies

$$\Delta W_{j+1} = W_{j+1}K \in o_{l,k,j+1,i,h} \quad \text{and} \quad \Delta W_j = -KW_j \in o_{lkjih}.$$

That is, each displacement in τ_{lkjih} is a sum of two one-matrix displacements in ϕ_{lkjih} and $\phi_{l,k,j+1,i,h}$. We think this is quite an elegant relationship, and it motivates why we define the two-matrix subspaces as we do. It will be very helpful in constructing a basis for the space tangent to the stratum at θ.

If all the one-matrix moves in ϕ_{lkjih} and $\phi_{l,k,j+1,i,h}$ stay on the stratum S that contains θ, then the two-matrix moves in τ_{lkjih} add nothing useful. But if the one-matrix moves leave the stratum, then the differential two-matrix move is interesting and useful, because its displacement $\Delta\theta$ is tangent to S. If, moreover, the one-matrix moves leave the nullspace of $\mathrm{d}\mu(\theta)$ (which implies that the moves leave the fiber), the differential two-matrix move is interesting because its displacement $\Delta\theta$ lies in $\operatorname{null} \mathrm{d}\mu(\theta)$. In the latter case, the two-matrix displacement satisfies $W_{L\sim j+1}\Delta W_{j+1}W_{j\sim 0} + W_{L\sim j}\Delta W_j W_{j-1\sim 0} = 0$ (recall the proof of Lemma 19), but those two terms are nonzero.

Recall that one-matrix moves in $\phi_{l,k,j+1,i,h}$ and ϕ_{lkjih} leave the fiber if and only if $l = L$ and $h = 0$. One-matrix moves in $\phi_{l,k,j+1,i,h}$ and ϕ_{lkjih} leave the stratum if they leave the fiber or if $l > k$ and $i > h$. (The latter condition implies a change in the rank of some subsequence matrix, i.e., a combinatorial move.) As every differential two-matrix move has $\Delta\theta \in \operatorname{null} \mathrm{d}\mu(\theta)$ (by Lemma 19) and $\Delta\theta$ tangent to the stratum S at θ (by Lemma 20), we are particularly interested in the differential two-matrix moves that combine two one-matrix moves that lack one or both of those properties, and we will use them to help form a prebasis for $\operatorname{null} \mathrm{d}\mu(\theta)$ and a prebasis for $T_\theta S$. Hence we define the sets

$$\begin{aligned}
\Theta_{\mathrm{T}}^{L0} &= \{\tau_{Lkji0} \in \Theta_{\mathrm{T}}\} = \{\tau_{Lkji0} \neq \{\mathbf{0}\} : L \geq k \geq j \geq i \geq 0 \text{ and } L > j > 0\} \quad \text{and} \\
\Theta_{\mathrm{T}}^{\mathrm{comb}} &= \{\tau_{lkjih} \in \Theta_{\mathrm{T}} : l > k \text{ and } i > h\} = \{\tau_{lkjih} \neq \{\mathbf{0}\} : L \geq l > k \geq j \geq i > h \geq 0\}.
\end{aligned}$$

Note that the notation $\Theta_{\mathrm{T}}^{\mathrm{comb}}$ is a bit of a misnomer, as the subspaces in $\Theta_{\mathrm{T}}^{\mathrm{comb}}$ do not represent combinatorial moves. (No finite two-matrix move changes the rank of any subsequence matrix.) Rather, they represent non-combinatorial replacements for the combinatorial subspaces in $\Theta_{\mathrm{O}}^{\mathrm{comb}}$.

6.6 The Fiber Prebasis and the Stratum Prebasis

The *fiber prebasis* at θ is a set of subspaces of $\mathbb{R}^{d_\theta}$ that spans null $\mathrm{d}\mu(\theta)$. Let S be the stratum of $\mu^{-1}(W)$ that contains θ. Let $T_\theta S$ be the space tangent to S at θ, and note that $T_\theta S \subseteq \operatorname{null} \mathrm{d}\mu(\theta)$. The *stratum prebasis* at θ is a set of subspaces of $\mathbb{R}^{d_\theta}$ that spans $T_\theta S$.

The fiber prebasis is

$$\begin{aligned}\Theta^{\mathrm{fiber}} &= \Theta_{\mathrm{O}}^{\mathrm{fiber}} \cup \Theta_{\mathrm{T}}^{L0} = \left(\Theta_{\mathrm{O}} \setminus \Theta_{\mathrm{O}}^{L0}\right) \cup \Theta_{\mathrm{T}}^{L0} \\ &= \left(\Theta_{\mathrm{O}} \setminus \{\phi_{lkjih} \in \Theta_{\mathrm{O}} : l = L \text{ and } h = 0\}\right) \cup \{\tau_{lkjih} \in \Theta_{\mathrm{T}} : l = L \text{ and } h = 0\}.\end{aligned}$$

The fiber prebasis includes every one-matrix subspace ϕ_{lkjih} such that $\phi_{lkjih} \subseteq \operatorname{null} \mathrm{d}\mu(\theta)$ and $\phi_{lkjih} \neq \{\mathbf{0}\}$, plus some additional two-matrix move subspaces as needed so that Θ^{fiber} spans null $\mathrm{d}\mu(\theta)$ (as we will show). The fiber prebasis excludes the one-matrix subspaces ϕ_{Lkji0} because they do not lie in null $\mathrm{d}\mu(\theta)$. (A one-matrix move with displacement $\Delta\theta \in \phi_{Lkji0} \setminus \{\mathbf{0}\}$ moves off the fiber—that is, it changes the value of $\mu(\theta)$.) By contrast, every two-matrix subspace satisfies $\tau_{lkjih} \subseteq \operatorname{null} \mathrm{d}\mu(\theta)$. (All differential two-matrix moves stay on null $\mathrm{d}\mu(\theta)$, and all finite two-matrix moves stay on the fiber and the stratum.)

Recall that a two-matrix subspace τ_{Lkji0} has the property that every displacement $\Delta\theta \in \tau_{Lkji0}$ is a linear combination of a displacement from $\phi_{L,k,j+1,i,0}$ and a displacement from ϕ_{Lkji0} (two subspaces we deliberately omit from Θ^{fiber}). Therefore, the two-matrix subspaces in Θ_{T}^{L0} have the property that $\operatorname{span} \Theta_{\mathrm{T}}^{L0} \subset \operatorname{span} \Theta_{\mathrm{O}}^{L0}$. However, $\operatorname{span} \Theta_{\mathrm{T}}^{L0}$ is a subset of null $\mathrm{d}\mu(\theta)$ whereas no subspace in Θ_{O}^{L0} is a subset of null $\mathrm{d}\mu(\theta)$. Observe that for each k and i satisfying $i, k \in [0, L]$ and $k \geq i - 1$, we are removing $k - i + 2$ one-matrix subspaces of dimension $\omega_{Li}\omega_{k0}$ and replacing them with $k - i + 1$ two-matrix subspaces of dimension $\omega_{Li}\,\omega_{k0}$.

The stratum prebasis is

$$\begin{aligned}\Theta^{\mathrm{stratum}} &= \left(\Theta_{\mathrm{O}} \setminus \Theta_{\mathrm{O}}^{L0} \setminus \Theta_{\mathrm{O}}^{\mathrm{comb}}\right) \cup \Theta_{\mathrm{T}}^{L0} \cup \Theta_{\mathrm{T}}^{\mathrm{comb}} \\ &= \left(\Theta_{\mathrm{O}} \setminus \{\phi_{lkjih} \in \Theta_{\mathrm{O}} : (l = L \text{ and } h = 0) \text{ or } (l > k \text{ and } i > h)\}\right) \cup \\ &\quad \{\tau_{lkjih} \in \Theta_{\mathrm{T}} : (l = L \text{ and } h = 0) \text{ or } (l > k \text{ and } i > h)\}.\end{aligned}$$

The stratum prebasis includes every one-matrix subspace $\phi_{lkjih} \neq \{\mathbf{0}\}$ that is tangent to S at θ, plus some additional two-matrix move subspaces as needed so that $\Theta^{\mathrm{stratum}}$ spans the tangent space $T_\theta S$. The stratum prebasis, like the fiber prebasis, excludes the one-matrix subspaces ϕ_{Lkji0} (because their moves move off the fiber), but it also excludes all the combinatorial moves (as combinatorial moves move off the stratum, though they stay on the fiber). The stratum prebasis, like the fiber prebasis, includes the two-matrix subspaces in Θ_{T}^{L0}, but it also includes the two-matrix subspaces in $\Theta_{\mathrm{T}}^{\mathrm{comb}}$. The latter have the property that $\operatorname{span} \Theta_{\mathrm{T}}^{\mathrm{comb}} \subset \operatorname{span} \Theta_{\mathrm{O}}^{\mathrm{comb}}$. However, $\operatorname{span} \Theta_{\mathrm{T}}^{\mathrm{comb}}$ and $\operatorname{span} \Theta_{\mathrm{T}}^{L0}$ are subsets of $T_\theta S$ whereas no subspace in $\operatorname{span} \Theta_{\mathrm{O}}^{\mathrm{comb}}$ nor $\operatorname{span} \Theta_{\mathrm{O}}^{L0}$ is a subset of $T_\theta S$. (Again, all differential two-matrix moves stay on null $\mathrm{d}\mu(\theta)$, and all finite two-matrix moves stay on the stratum.)

The addition of the subspaces in $\Theta_{\mathrm{T}}^{\mathrm{comb}}$ deserves more explanation. Every subspace in $\Theta_{\mathrm{T}}^{\mathrm{comb}}$ is related to swapping moves (not connecting moves). Recall that a swapping move comes from a subspace ϕ_{lkjih} where $l > k \geq i > h$. Given fixed values of l, k, i, and h, let S' be the stratum for which ω_{li} and ω_{kh} are one less and

ω_{lh} and ω_{ki} are one greater than they are for θ and the other points on S. The subspaces ϕ_{lkjih} with $j \in [i, k+1]$ represent one-matrix moves that all move from θ into the stratum S' (by increasing the rank of one or more subsequence matrices). There are $k - i + 2$ such subspaces, which we omit from Θ^{stratum}. However, for each $j \in [i, k]$, there is a two-matrix subspace $\tau_{lkjih} \subseteq T_\theta S$ whose members are linear combinations of one-matrix displacements in ϕ_{lkjih} and $\phi_{l,k,j+1,i,h}$; we include it in Θ^{stratum}. Observe that for each l, k, i, and h satisfying $L \geq l > k \geq i > h \geq 0$, we are removing $k - i + 2$ one-matrix subspaces of dimension $\omega_{li}\omega_{kh}$ and replacing them with $k - i + 1$ two-matrix subspaces of dimension $\omega_{li}\,\omega_{kh}$.

The rest of this section is devoted to showing that Θ^{fiber} is a prebasis for $\operatorname{null} \mathrm{d}\mu(\theta)$ and Θ^{stratum} is a prebasis for $T_\theta S$. We know that every subspace in Θ^{fiber} is a subset of $\operatorname{null} \mathrm{d}\mu(\theta)$ and every subspace in Θ^{stratum} is a subset of $T_\theta S$. As a second step, Lemma 22 below shows that the subspaces in Θ^{fiber} are linearly independent, and Lemma 24 shows that the subspaces in Θ^{stratum} are linearly independent. As a third step, we will add up the dimensions of the subspaces in Θ^{fiber} and see that the total dimension is the dimension of $\operatorname{null} \mathrm{d}\mu$. Likewise, the sum of the dimensions of the subspaces in Θ^{stratum} is the dimension of $T_\theta S$. These three steps together deliver the desired results.

Lemma 22. *The subspaces in* Θ^{fiber} *are linearly independent.*

Proof. Suppose for the sake of contradiction that Θ^{fiber} is not linearly independent. Then we can choose one vector from each subspace in Θ^{fiber}—call them *canceling vectors*—such that the sum of all the canceling vectors is zero, and at least two canceling vectors are nonzero. By Lemma 11, the one-matrix move subspaces $\phi_{lkjih} \in \Theta_{\text{O}}$ are linearly independent. Therefore, at least one canceling vector from a two-matrix subspace in Θ_{T}^{L0} is nonzero. Let $v \in \tau_{Lkji0}$ be the two-matrix canceling vector with minimum index j such that $v \neq \mathbf{0}$. Let V_j be the matrix in the W_j position of v. Then $V_j \in o_{Lkji0}$. By Lemma 21, $V_j \neq 0$ (as $v \neq \mathbf{0}$).

Let X_j be the matrix in the W_j position of the sum of all the canceling vectors; by assumption, $X_j = 0$. Then X_j is a sum of V_j and contributions from other canceling vectors. However, as $\phi_{Lkji0} \notin \Theta^{\text{fiber}}$, those other contributions are linearly independent of V_j, with the possible exception of a contribution from a canceling vector $v' \in \tau_{L,k,j-1,i,0}$ (if $j > i$; otherwise $\tau_{L,k,j-1,i,0}$ is not defined). However, we assumed that $v \in \tau_{Lkji0}$ is the two-matrix canceling vector with minimum index j such that $v \neq \mathbf{0}$; so even if $j > i$, we have $v' = \mathbf{0}$. It follows that X_j is not zero, contradicting the assumption that the sum of the canceling vectors is zero. From this contradiction, it follows that the subspaces in Θ^{fiber} are linearly independent. □

Theorem 23. Θ^{fiber} *is a prebasis for* $\operatorname{null} \mathrm{d}\mu(\theta)$. *In particular,* $\operatorname{null} \mathrm{d}\mu(\theta) = \operatorname{span} \Theta^{\text{fiber}}$.

Proof. By Lemma 22, the subspaces in Θ^{fiber} are linearly independent. Therefore, the dimension of the space spanned by Θ^{fiber} is equal to the sum of the dimensions of all the subspaces in Θ^{fiber}. In Section 6.7, we will show that this sum is

$$D^{\text{fiber}} = d_\theta - \sum_{j=1}^{L} \operatorname{rk} W_{L\sim j} \cdot \operatorname{rk} W_{j-1\sim 0} + \sum_{j=1}^{L-1} \operatorname{rk} W_{L\sim j} \cdot \operatorname{rk} W_{j\sim 0}.$$

Corollary 35 in Appendix A shows that $\operatorname{null} \mathrm{d}\mu(\theta)$ has the same dimension; that is, $\dim \operatorname{null} \mathrm{d}\mu(\theta) = D^{\text{fiber}}$.

By Lemmas 18 and 19, each subspace in Θ^{fiber} is a subspace of $\operatorname{null} \mathrm{d}\mu(\theta)$, so $\operatorname{span} \Theta^{\text{fiber}} \subseteq \operatorname{null} \mathrm{d}\mu(\theta)$. As the two spaces have the same dimension, $\operatorname{span} \Theta^{\text{fiber}} = \operatorname{null} \mathrm{d}\mu(\theta)$. Therefore, Θ^{fiber} is a prebasis for $\operatorname{null} \mathrm{d}\mu(\theta)$. □

Lemma 24. *The subspaces in* Θ^{stratum} *are linearly independent.*

Proof. Essentially the same as the proof of Lemma 22, with the following changes. Θ^{stratum} has fewer one-matrix subspaces and more two-matrix subspaces related to the omitted one-matrix subspaces; but as in the proof of Lemma 22, observe that each omitted one-matrix subspace ϕ_{lkjih} can receive contributions from at most two one-matrix subspaces, τ_{lkjih} (if $k \geq j$) and $\tau_{l,k,j-1,i,h}$ (if $j > i$). Define the canceling vectors the same way, and let $v \in \tau_{lkjih}$ be the two-matrix canceling vector with minimum index j such that $v \neq \mathbf{0}$ (so that the canceling vector from $\tau_{l,k,j-1,i,h}$ is zero or $\tau_{l,k,j-1,i,h}$ does not exist.) The logic of the proof is unchanged: v makes a nonzero contribution to the subspace spanned by ϕ_{lkjih} which is not canceled by any other subspace, contradicting the assumption that there exist canceling vectors that are not all zero. □

Theorem 25. Θ^{stratum} *is a prebasis for* $T_\theta S$. *In particular,* $T_\theta S = \operatorname{span} \Theta^{\text{stratum}}$.

Proof. By Lemma 24, the subspaces in Θ^{stratum} are linearly independent. Therefore, the dimension of the space spanned by Θ^{stratum} is equal to the sum of the dimensions of all the subspaces in Θ^{stratum}. In Section 6.7, we will show that this sum is

$$D^{\text{stratum}} = d_\theta - \omega_{L0}(d_L + d_0 - \omega_{L0}) - \sum_{L \geq k+1 \geq i > 0} \beta_{k+1,i,i}\, \alpha_{k,k,i-1}.$$

By Lemma 20, each subspace in Θ^{stratum} is a subspace of $T_\theta S$, so $\operatorname{span} \Theta^{\text{stratum}} \subseteq T_\theta S$. Hence the dimension of $T_\theta S$ is at least D^{stratum}. However, let $N_\theta S \subset \mathbb{R}^{d_\theta}$ be the subspace normal to S at θ; we can also show that the dimension of $N_\theta S$ is at least

$$\omega_{L0}(d_L + d_0 - \omega_{L0}) + \sum_{L \geq k+1 \geq i > 0} \beta_{k+1,i,i}\, \alpha_{k,k,i-1}.$$

The sum of these lower bounds on the dimensions of $T_\theta S$ and $N_\theta S$ is d_θ, so both bounds must be tight. Hence, $\dim T_\theta S = D^{\text{stratum}}$ and $\operatorname{span} \Theta^{\text{stratum}} = T_\theta S$. Therefore, Θ^{stratum} is a prebasis for $T_\theta S$. □

6.7 Counting the Degrees of Freedom

Let D_{O}, D_{O}^{L0}, $D_{\text{O}}^{\text{fiber}}$, $D_{\text{O}}^{\text{comb}}$, D_{T}^{L0}, $D_{\text{T}}^{\text{comb}}$, D^{fiber}, and D^{stratum} (and so forth) denote the dimension of the subspace spanned by the prebasis Θ_{O}, Θ_{O}^{L0}, $\Theta_{\text{O}}^{\text{fiber}}$, $\Theta_{\text{O}}^{\text{comb}}$, Θ_{T}^{L0}, $\Theta_{\text{T}}^{\text{comb}}$, Θ^{fiber}, and Θ^{stratum}, respectively. Then $D_{\text{O}} = d_\theta$, as the one-matrix move prebasis Θ_{O} spans the entire weight space $\mathbb{R}^{d_\theta}$ by Lemma 11. Table 5 gives the definitions of several prebases and the dimensions of the subspaces (of $\mathbb{R}^{d_\theta}$) spanned by those prebases. In this section we derive those dimensions. (See Appendix B for some additional prebases.)

These numbers tell us the number of degrees of freedom of motion from a point $\theta \in \mu^{-1}(W)$ that have certain properties—for instance, the degrees of freedom of one-matrix moves that stay on the fiber, one-matrix moves that move off the fiber, or moves (one- and two-matrix) that are tangent to the stratum. Some of these counts are needed to prove that Θ^{fiber} spans $\operatorname{null} \mathrm{d}\mu(\theta)$ and Θ^{stratum} spans the space tangent to the stratum containing θ. However, this section makes for mind-numbing reading and can be safely skipped. We provide it as a reference for anyone who needs to know the dimensions of specific subspaces, who wishes to check our proofs carefully, or who wishes to extend results and ideas in this paper.

Recall that a one-matrix subspace $\phi_{lkjih} \in \Theta_{\text{O}}$ or a two-matrix subspace $\tau_{lkjih} \in \Theta_{\text{T}}$ has dimension $\omega_{li}\, \omega_{kh}$. If the subspaces in a set (such as Θ^{stratum}) are linearly independent, then the dimension of the space they span is equal to the sum of the dimensions of the subspaces in the set. Therefore, if Θ' is such a set, the dimension of the space Θ' spans is

$$D' = \sum_{\phi_{lkjih} \in \Theta'} \omega_{li}\, \omega_{kh} + \sum_{\tau_{lkjih} \in \Theta'} \omega_{li}\, \omega_{kh}.$$

$$\Theta_{\mathrm{O}} = \{\phi_{lkjih} \neq \{\mathbf{0}\} : L \geq l \geq j \geq i \geq 0 \text{ and } L \geq k \geq j-1 \geq h \geq 0\}$$

$$D_{\mathrm{O}} = d_\theta = \sum_{j=1}^{L} d_j d_{j-1} \tag{11}$$

$$\Theta_{\mathrm{O}}^{L0} = \{\phi_{Lkji0} \in \Theta_{\mathrm{O}}\} = \{\phi_{Lkji0} \neq \{\mathbf{0}\} : L \geq k \geq j-1 \geq 0 \text{ and } L \geq j \geq i \geq 0\}$$

$$D_{\mathrm{O}}^{L0} = \sum_{j=1}^{L} \left(\sum_{i=0}^{j} \omega_{Li} \right) \left(\sum_{k=j-1}^{L} \omega_{k0} \right) = \sum_{j=1}^{L} \operatorname{rk} W_{L\sim j} \cdot \operatorname{rk} W_{j-1\sim 0} \tag{12}$$

$$\Theta_{\mathrm{O}}^{\mathrm{fiber}} = \Theta_{\mathrm{O}} \setminus \Theta_{\mathrm{O}}^{L0} = \{\phi_{lkjih} \in \Theta_{\mathrm{O}} : L > l \text{ or } h > 0\}$$

$$D_{\mathrm{O}}^{\mathrm{fiber}} = D_{\mathrm{O}} - D_{\mathrm{O}}^{L0} = d_\theta - \sum_{j=1}^{L} \operatorname{rk} W_{L\sim j} \cdot \operatorname{rk} W_{j-1\sim 0} \tag{13}$$

$$\Theta_{\mathrm{O}}^{\mathrm{comb}} = \{\phi_{lkjih} \in \Theta_{\mathrm{O}} : l > k \text{ and } i > h\} = \{\phi_{lkjih} \neq \{\mathbf{0}\} : L \geq l \geq k+1 \geq j \geq i > h \geq 0\}$$

$$D_{\mathrm{O}}^{\mathrm{comb}} = \sum_{L \geq k+1 \geq i > 0} (k-i+2) \underbrace{(\operatorname{rk} W_{k+1\sim i} - \operatorname{rk} W_{k+1\sim i-1})}_{\beta_{k+1,i,i}} \underbrace{(\operatorname{rk} W_{k\sim i-1} - \operatorname{rk} W_{k+1\sim i-1})}_{\alpha_{k,k,i-1}} \tag{14}$$

$$\Theta_{\mathrm{O}}^{L0,\neg\mathrm{comb}} = \Theta_{\mathrm{O}}^{L0} \setminus \Theta_{\mathrm{O}}^{\mathrm{comb}} = \{\phi_{Lkji0} \in \Theta_{\mathrm{O}} : L = k \text{ or } i = 0\}$$

$$D_{\mathrm{O}}^{L0,\neg\mathrm{comb}} = \omega_{L0} \left(\sum_{i=1}^{L} (L-i+1)\, \omega_{Li}\, \omega_{L0} + \sum_{k=0}^{L-1} (k+1)\, \omega_{L0}\, \omega_{k0} + L\, \omega_{L0} \right) \tag{15}$$

$$\Theta_{\mathrm{T}} = \{\tau_{lkjih} \neq \{\mathbf{0}\} : L \geq l > j \geq i \geq 0 \text{ and } L \geq k \geq j > h \geq 0\}$$

$$\Theta_{\mathrm{T}}^{L0} = \{\tau_{Lkji0} \in \Theta_{\mathrm{T}}\} = \{\tau_{Lkji0} \neq \{\mathbf{0}\} : L \geq k \geq j \geq i \geq 0 \text{ and } L > j > 0\}$$

$$D_{\mathrm{T}}^{L0} = \sum_{j=1}^{L-1} \left(\sum_{i=0}^{j} \omega_{Li} \right) \left(\sum_{k=j}^{L} \omega_{k0} \right) = \sum_{j=1}^{L-1} \operatorname{rk} W_{L\sim j} \cdot \operatorname{rk} W_{j\sim 0} \tag{16}$$

$$\Theta_{\mathrm{T}}^{\mathrm{comb}} = \{\tau_{lkjih} \in \Theta_{\mathrm{T}} : l > k \text{ and } i > h\} = \{\tau_{lkjih} \neq \{\mathbf{0}\} : L \geq l > k \geq j \geq i > h \geq 0\}$$

$$D_{\mathrm{T}}^{\mathrm{comb}} = \sum_{L > k \geq i > 0} (k-i+1) \underbrace{(\operatorname{rk} W_{k+1\sim i} - \operatorname{rk} W_{k+1\sim i-1})}_{\beta_{k+1,i,i}} \underbrace{(\operatorname{rk} W_{k\sim i-1} - \operatorname{rk} W_{k+1\sim i-1})}_{\alpha_{k,k,i-1}} \tag{17}$$

$$\Theta_{\mathrm{T}}^{L0,\neg\mathrm{comb}} = \Theta_{\mathrm{T}}^{L0} \setminus \Theta_{\mathrm{T}}^{\mathrm{comb}} = \{\tau_{Lkji0} \in \Theta_{\mathrm{T}} : L = k \text{ or } i = 0\}$$

$$D_{\mathrm{T}}^{L0,\neg\mathrm{comb}} = \omega_{L0} \left(\sum_{i=1}^{L-1} (L-i)\, \omega_{Li}\, \omega_{L0} + \sum_{k=1}^{L-1} k\, \omega_{L0}\, \omega_{k0} + (L-1)\, \omega_{L0} \right) \tag{18}$$

$$\Theta^{\mathrm{fiber}} = \Theta_{\mathrm{O}}^{\mathrm{fiber}} \cup \Theta_{\mathrm{T}}^{L0} = \left(\Theta_{\mathrm{O}} \setminus \Theta_{\mathrm{O}}^{L0} \right) \cup \Theta_{\mathrm{T}}^{L0}$$

$$D^{\mathrm{fiber}} = D_{\mathrm{O}} - D_{\mathrm{O}}^{L0} + D_{\mathrm{T}}^{L0} = d_\theta - \sum_{j=1}^{L} \operatorname{rk} W_{L\sim j} \cdot \operatorname{rk} W_{j-1\sim 0} + \sum_{j=1}^{L-1} \operatorname{rk} W_{L\sim j} \cdot \operatorname{rk} W_{j\sim 0} \tag{19}$$

$$\Theta^{\mathrm{stratum}} = \left(\Theta_{\mathrm{O}} \setminus \Theta_{\mathrm{O}}^{\mathrm{comb}} \setminus \Theta_{\mathrm{O}}^{L0} \right) \cup \Theta_{\mathrm{T}}^{\mathrm{comb}} \cup \Theta_{\mathrm{T}}^{L0}$$

$$D^{\mathrm{stratum}} = d_\theta - \operatorname{rk} W\, (d_L + d_0 - \operatorname{rk} W) - \sum_{L \geq k+1 \geq i > 0} \underbrace{(\operatorname{rk} W_{k+1\sim i} - \operatorname{rk} W_{k+1\sim i-1})}_{\beta_{k+1,i,i}} \underbrace{(\operatorname{rk} W_{k\sim i-1} - \operatorname{rk} W_{k+1\sim i-1})}_{\alpha_{k,k,i-1}} \tag{20}$$

Table 5: Sets of subspaces of moves and their total degrees of freedom. See also Table 6.

As Θ_{O} is a prebasis by Lemma 11, its members (the one-matrix subspaces) are linearly independent, so we can apply this formula to any subset of Θ_{O}.

In Section 6.1 we defined $\Theta_{\mathrm{O}}^{L0} \subseteq \Theta_{\mathrm{O}}$, representing the one-matrix moves that move off the fiber (change $\mu(\theta)$), as a prelude to defining $\Theta_{\mathrm{O}}^{\mathrm{fiber}} = \Theta_{\mathrm{O}} \setminus \Theta_{\mathrm{O}}^{L0}$, representing the one-matrix moves that stay on the fiber (don't change $\mu(\theta)$). Θ_{O}^{L0} contains all the one-matrix subspaces ϕ_{lkjih} such that $l = L$ and $h = 0$, and $\Theta_{\mathrm{O}}^{\mathrm{fiber}}$ contains all the other one-matrix subspaces. From the definition $\Theta_{\mathrm{O}}^{L0} = \{\phi_{Lkji0} \neq \{\mathbf{0}\} : L \geq k \geq j-1 \geq 0 \text{ and } L \geq j \geq i \geq 0\}$, we obtain the first formula of (12) for D_{O}^{L0}, the dimension of the space spanned by Θ_{O}^{L0} (see Table 5). The second formula of (12) follows from the identity (7).

In Section 6.5 we defined $\Theta_{\mathrm{T}}^{L0} \subseteq \Theta_{\mathrm{T}}$, which contains all the two-matrix subspaces τ_{lkjih} such that $l = L$ and $h = 0$. The subspaces in Θ_{T}^{L0} are linearly independent by Lemma 22, so the dimension D_{T}^{L0} of the space spanned by Θ_{T}^{L0} is equal to the sum of the dimensions of the subspaces in Θ_{T}^{L0}. From the definition $\Theta_{\mathrm{T}}^{L0} = \{\tau_{Lkji0} \neq \{\mathbf{0}\} : L \geq k \geq j \geq i \geq 0 \text{ and } L > j > 0\}$ and the identity (7), we obtain the formulae (16) for D_{T}^{L0} (see Table 5).

The space spanned by $\Theta_{\mathrm{O}}^{\mathrm{fiber}} = \Theta_{\mathrm{O}} \setminus \Theta_{\mathrm{O}}^{L0}$ has dimension $D_{\mathrm{O}}^{\mathrm{fiber}} = D_{\mathrm{O}} - D_{\mathrm{O}}^{L0}$; see the formula (13). Note that if we consider only moves in $\Theta_{\mathrm{O}}^{\mathrm{fiber}}$ that change a specific factor matrix W_j, the space spanned by those moves has dimension $D_j^{\mathrm{fiber}} = d_j d_{j-1} - \operatorname{rk} W_{L\sim j} \cdot \operatorname{rk} W_{j-1\sim 0}$, which not coincidentally is the dimension of N_j (defined in Section 6.1).

A particularly important prebasis is $\Theta^{\mathrm{fiber}} = \Theta_{\mathrm{O}}^{\mathrm{fiber}} \cup \Theta_{\mathrm{T}}^{L0}$, which spans $\operatorname{null} \mathrm{d}\mu(\theta)$ by Theorem 23. By Lemma 22, the subspaces in $\Theta_{\mathrm{O}}^{\mathrm{fiber}}$ and in Θ_{T}^{L0} are (separately and together) linearly independent. Hence, $D^{\mathrm{fiber}} = D_{\mathrm{O}}^{\mathrm{fiber}} + D_{\mathrm{T}}^{L0}$, giving us the formula (19). That formula is the same as the dimension of $\operatorname{null} \mathrm{d}\mu(\theta)$ according to Corollary 35—which is how we know that Θ^{fiber} spans all of $\operatorname{null} \mathrm{d}\mu(\theta)$.

Recall from Section 6.3 that the infinitesimal combinatorial moves are the one-matrix moves with a sufficiently small displacement $\Delta\theta \in \phi_{lkjih} \setminus \{\mathbf{0}\}$ where $l > k$ and $i > h$. The set containing these subspaces is $\Theta_{\mathrm{O}}^{\mathrm{comb}} = \{\phi_{lkjih} \neq \{\mathbf{0}\} : L \geq l \geq k+1 \geq j \geq i > h \geq 0\}$. To derive the dimension of the space spanned by $\Theta_{\mathrm{O}}^{\mathrm{comb}}$, we use the identities (6) and (7) and the fact that in the first summation below, the term $\omega_{li}\,\omega_{kh}$ appears $k - i + 2$ times—once for each $j \in [i, k+1]$.

$$
\begin{aligned}
D_{\mathrm{O}}^{\mathrm{comb}} &= \sum_{L \geq l \geq k+1 \geq j \geq i > h \geq 0} \omega_{li}\,\omega_{kh} = \sum_{L \geq l \geq k+1 \geq i > h \geq 0} (k-i+2)\,\omega_{li}\,\omega_{kh} \\
&= \sum_{L \geq k+1 \geq i > 0} (k-i+2) \left(\sum_{l=k+1}^{L} \omega_{li} \right) \left(\sum_{h=0}^{i-1} \omega_{kh} \right) = \sum_{L \geq k+1 \geq i > 0} (k-i+2) \beta_{k+1,i,i}\, \alpha_{k,k,i-1} \\
&= \sum_{L \geq k+1 \geq i > 0} (k-i+2)\,(\operatorname{rk} W_{k+1\sim i} - \operatorname{rk} W_{k+1\sim i-1})\,(\operatorname{rk} W_{k\sim i-1} - \operatorname{rk} W_{k+1\sim i-1}).
\end{aligned}
$$

It follows from Lemma 24 that the subspaces in $\Theta_{\mathrm{T}}^{\mathrm{comb}}$ are linearly independent. The process of determining $D_{\mathrm{T}}^{\mathrm{comb}}$ is nearly the same as for $D_{\mathrm{O}}^{\mathrm{comb}}$, but in $\Theta_{\mathrm{T}}^{\mathrm{comb}} = \{\tau_{lkjih} \neq \{\mathbf{0}\} : L \geq l > k \geq j \geq i > h \geq 0\}$ the indices have the constraint $k \geq j$ (rather than $k \geq j-1$). Hence the term $\omega_{li}\,\omega_{kh}$ appears once for each $j \in [i, k]$ and

$$
D_{\mathrm{T}}^{\mathrm{comb}} = \sum_{L \geq l > k \geq j \geq i > h \geq 0} \omega_{li}\,\omega_{kh} = \sum_{L \geq l > k \geq i > h \geq 0} (k-i+1)\,\omega_{li}\,\omega_{kh} = \sum_{L > k \geq i > 0} (k-i+1) \beta_{k+1,i,i}\, \alpha_{k,k,i-1}.
$$

When we derive the dimension D^{stratum} of the stratum that contains θ, we will use the difference between $D_{\mathrm{O}}^{\mathrm{comb}}$ and $D_{\mathrm{T}}^{\mathrm{comb}}$, which is

$$
D_{\mathrm{O}}^{\mathrm{comb}} - D_{\mathrm{T}}^{\mathrm{comb}} = \sum_{L \geq k+1 \geq i > 0} \beta_{k+1,i,i}\, \alpha_{k,k,i-1}.
$$

Let $\Theta_{\mathrm{O}}^{L0,\neg\mathrm{comb}} = \Theta_{\mathrm{O}}^{L0} \setminus \Theta_{\mathrm{O}}^{\mathrm{comb}}$, a prebasis for the moves that move off the fiber but are not combinatorial; that is, they change $\mu(\theta)$ but do not change the rank of any subsequence matrix. (We will use $\Theta_{\mathrm{O}}^{L0,\neg\mathrm{comb}}$ later to derive D^{stratum}.) The set $\Theta_{\mathrm{O}}^{L0,\neg\mathrm{comb}}$ contains every one-matrix subspace ϕ_{Lkji0} such that $k = L$ or $i = 0$. The dimension of the space spanned by $\Theta_{\mathrm{O}}^{L0,\neg\mathrm{comb}}$ is

$$\begin{aligned} D_{\mathrm{O}}^{L0,\neg\mathrm{comb}} &= \sum_{L=k\geq j\geq i>0 \text{ or } L\geq k+1\geq j>i=0 \text{ or } L=k\geq j>i=0} \omega_{Li}\,\omega_{k0} \\ &= \sum_{i=1}^{L}(L-i+1)\,\omega_{Li}\,\omega_{L0} + \sum_{k=0}^{L-1}(k+1)\,\omega_{L0}\,\omega_{k0} + L\,\omega_{L0}^2, \end{aligned}$$

giving us the formula (15). Analogously, let $\Theta_{\mathrm{T}}^{L0,\neg\mathrm{comb}} = \Theta_{\mathrm{T}}^{L0} \setminus \Theta_{\mathrm{T}}^{\mathrm{comb}}$. This set contains every two-matrix subspace τ_{Lkji0} such that $k = L$ or $i = 0$. The dimension of the space spanned by $\Theta_{\mathrm{T}}^{L0,\neg\mathrm{comb}}$ is

$$\begin{aligned} D_{\mathrm{T}}^{L0,\neg\mathrm{comb}} &= \sum_{L=k>j\geq i>0 \text{ or } L>k\geq j>i=0 \text{ or } L=k>j>i=0} \omega_{Li}\,\omega_{k0} \\ &= \sum_{i=1}^{L-1}(L-i)\,\omega_{Li}\,\omega_{L0} + \sum_{k=1}^{L-1} k\,\omega_{L0}\,\omega_{k0} + (L-1)\,\omega_{L0}^2, \end{aligned}$$

giving us the formula (18). Our derivation of D^{stratum} will use the difference between $D_{\mathrm{O}}^{L0,\neg\mathrm{comb}}$ and $D_{\mathrm{T}}^{L0,\neg\mathrm{comb}}$. With the help of identity (1), we can simplify this difference to

$$\begin{aligned} D_{\mathrm{O}}^{L0,\neg\mathrm{comb}} - D_{\mathrm{T}}^{L0,\neg\mathrm{comb}} &= \sum_{i=1}^{L}\omega_{Li}\,\omega_{L0} + \sum_{k=0}^{L-1}\omega_{L0}\,\omega_{k0} + \omega_{L0}^2 \\ &= \omega_{L0}\left(\sum_{i=0}^{L}\omega_{Li} + \sum_{k=0}^{L}\omega_{k0} - \omega_{L0}\right) \\ &= \omega_{L0}(d_L + d_0 - \omega_{L0}). \end{aligned}$$

We now derive the dimension D^{stratum} of the space spanned by the stratum prebasis $\Theta^{\mathrm{stratum}}$, which is also the dimension of the stratum that contains θ.

$$\begin{aligned} D^{\mathrm{stratum}} &= \dim \Theta^{\mathrm{stratum}} \\ &= \dim\left(\left(\Theta_{\mathrm{O}} \setminus \Theta_{\mathrm{O}}^{\mathrm{comb}} \setminus \Theta_{\mathrm{O}}^{L0}\right) \cup \Theta_{\mathrm{T}}^{\mathrm{comb}} \cup \Theta_{\mathrm{T}}^{L0}\right) \\ &= \dim\left(\left(\Theta_{\mathrm{O}} \setminus \Theta_{\mathrm{O}}^{\mathrm{comb}} \setminus \Theta_{\mathrm{O}}^{L0,\neg\mathrm{comb}}\right) \cup \Theta_{\mathrm{T}}^{\mathrm{comb}} \cup \Theta_{\mathrm{T}}^{L0,\neg\mathrm{comb}}\right) \\ &= D_{\mathrm{O}} - D_{\mathrm{O}}^{\mathrm{comb}} - D_{\mathrm{O}}^{L0,\neg\mathrm{comb}} + D_{\mathrm{T}}^{\mathrm{comb}} + D_{\mathrm{T}}^{L0,\neg\mathrm{comb}} \\ &= D_{\mathrm{O}} - \left(D_{\mathrm{O}}^{L0,\neg\mathrm{comb}} - D_{\mathrm{T}}^{L0,\neg\mathrm{comb}}\right) - \left(D_{\mathrm{O}}^{\mathrm{comb}} - D_{\mathrm{T}}^{\mathrm{comb}}\right) \\ &= d_\theta - \omega_{L0}(d_L + d_0 - \omega_{L0}) - \sum_{L\geq k+1\geq i>0} \beta_{k+1,i,i}\,\alpha_{k,k,i-1}, \end{aligned}$$

giving us the formula (20).

6.8 The Hierarchy of Strata

In order to understand the topology of the fiber, it is necessary to understand the connectivity of the strata that comprise it. Consider two strata S_A and S_B with interval diagrams A and B respectively. We already know that if there exists a sequence of infinitesimal combinatorial moves from S_B to S_A, then S_B is in the closure of S_A, as shown simply in Lemma 26. However, in order to completely characterize the fiber, we must investigate whether it is always true that whenever S_B is in the closure of S_A, that there exists a sequence of infinitesimal combinatorial moves (swapping or connecting moves) that convert B into A. As we will see in Algorithm 2 and Corollary 33, the answer to this question is yes.

Lemma 26. *Consider two strata S_A and S_B with corresponding interval diagrams A and B. If there exists a sequence of infinitesimal combinatorial moves from S_B to S_A, then S_B is in the closure of S_A.*

Proof. Performing a combinatorial move at a stratum S can either cause us to remain on S or to move to a stratum S' such that S is in the closure of S', but never the other way. Hence by induction on the sequence of combinatorial moves, and the transitivity of closure, we conclude that if there is a sequence of infinitesimal combinatorial moves from S_B to S_A then S_B must be in the closure of S_A. □

First we define a relation $A \geq_r B$ between interval diagrams which is equivalent to the condition that the corresponding stratum S_B is in the closure of S_A. For two interval diagrams A and B of the same neural network, we say that $A \geq_r B$ if the rank of every subsequence product matrix in A is at least as much as the rank of the corresponding subsequence product matrix in B. Furthermore we define the relationship $A \geq_s B$ on two interval diagrams satisfying $A \geq_r B$ if an additional constraint is met. We sort all the interval lengths of A and B. If one of the two lists of interval lengths is shorter than the other, then it is padded with zeros. Then if every interval length of B can be matched with a higher interval length of A, then we say that $A \geq_s B$. Figure 10 shows an example of interval diagrams A and B satisfying both $A \geq_s B$ and $A \geq_r B$. Figure 11 shows an example of interval diagrams where $A \geq_r B$ is satisfied, but $A \geq_s B$ is not satisfied. Figure 12 provides an example of interval diagrams satisfying neither relation. Furthermore, swapping A and B in Figure 10 and Figure 11 give two additional examples of interval diagrams not satisfying either relation.

Figure 10: $A \geq_s B$. Sorting the lengths of the intervals of B gives $(3, 3, 2, 1)$, and sorting the lengths of the intervals of A gives $(4, 3, 3, 2)$. Since $4 \geq 3$, $3 \geq 3$, $3 \geq 2$, and $2 \geq 1$, we conclude that $A \geq_s B$.

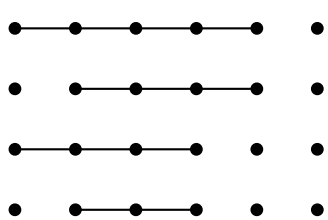


Figure 10: Interval diagram A.

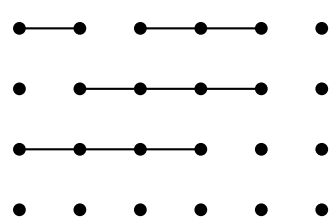


Figure 10: Interval diagram B.

Lemma 27. *(Weak Interval Matching Lemma) Consider two interval diagrams A and B of the same neural network satisfying $A \geq_r B$. If A and B additionally satisfy $A \geq_s B$, then every interval of B can be matched with a unique sub-interval interval of A, such that the matching function is injective.*

Proof. It is sufficient to show that Algorithm 1 produces an injective function M from the intervals of B to the intervals of A. Assume for contradiction that the interval $k \sim l$ of B was not matched to any sub-interval of A. This means that the longest interval of A starting at layer l is of length less than $k - l$. But then the

Figure 11: $A \geq_r B$ but $A \ngeq_s B$. Sorting the lengths of the intervals of B gives $(3, 2, 1)$, and sorting the lengths of the intervals of A gives $(4, 2, 0)$, so $A \ngeq_s B$. The reader can verify that every sub-interval in B exists in A.

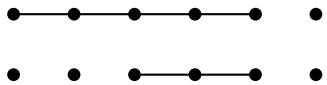


Figure 11: Interval diagram A.

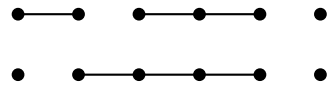


Figure 11: Interval diagram B.

Figure 12: $A \ngeq_r B$. Observe that in A, $\omega_{2\sim1} = 1$, but in B, $\omega_{2\sim1} = 2$, meaning that matrix rank was lost in going from B to A. Thus $A \ngeq_r B$.

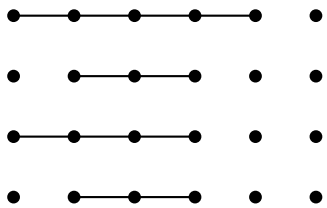


Figure 12: Interval diagram A.

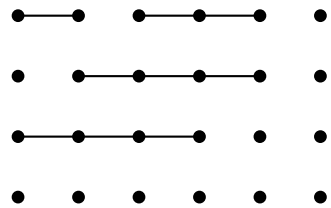


Figure 12: Interval diagram B.

rank of $W_{k\sim l}$ must be lower in A than it is in B, a contradiction, completing the proof. One way to see that this will never happen is revealed by analyzing the sorted rank lists $L_A = (a_1, a_2, \ldots)$ and $L_B = (b_1, b_2, \ldots)$ of A and B respectively. Since $A \geq_s B$, $a_i \geq b_i$ for every i. Then at every step of the innermost for-loop in Algorithm 1, some a_i and its corresponding b_i get reduced by the same number simultaneously. The former happens as unit intervals are removed from A, whereas the latter happens as the corresponding unit intervals in B are visited. Hence Algorithm 1 cannot result in a state where for some i, $a_i < b_i$, meaning that for every i, $a_i = 0$ only if $b_i = 0$, meaning that the algorithm cannot terminate with unmatched intervals in B.

Algorithm 1: Interval Matching

```
begin
    M ← Initialize mapping from intervals of B to sub-intervals of A
    for layer l ← 0 to d do
        I_{B,l} ← Intervals of B starting at layer l sorted from longest to shortest
        for interval (k ~ l)_B ∈ I_B do
            I_{A,k,l} ← Sub-intervals of A that are supersets of (k ~ l)_A in A
            Sort I_{A,k,l} ascending, on interval length
            for sub-interval (k' ~ l')_A ∈ I_{A,k,l} do
                Insert mapping (k ~ l)_B ↦ (k ~ l)_A into M
                Remove from A (and thus I_{A,k,l}) one multiplicity of all the unit intervals between k
                  and l.
                if k < d then Remove from A one multiplicity of k + 1 ~ k
                if l > 0 then Remove from A one multiplicity of l ~ l − 1
            end
        end
    end
    return M
end
```

□

Corollary 28. *Using the matchings from Lemma 27, we can perform a series of connecting moves to reach diagram A from B if $A \geq_s B$.*

Proof. Iteratively connect two intervals $k \sim l$ and $k' \sim l'$ of B which are mapped to the same interval a in A, when there is no interval $k'' \sim l''$ in B which is mapped to a and k'' and l'' both lie between $k \sim l$ and $k' \sim l'$. If $k \sim l$ is the only interval mapped to a, then extend it on both sides until $k \sim l$ grows to the same size as a. □

Thus we have a relatively intuitive algorithm in the special case of $A \geq_s B$. The ability to move from one stratum to another is rather useful both for applications in escaping spurious critical points as well as for gaining a better theoretical understanding of the topology of the fiber. Hence we would like to extend these results to the more general case of $A \geq_r B$. First we prove a useful lemma about interval multiplicities and matrix ranks.

Lemma 29. *Let $L \geq l > k \geq i > h \geq 0$. Then,* $\operatorname{rk} W_{l\sim h} + \operatorname{rk} W_{k\sim i} - \operatorname{rk} W_{l\sim i} - \operatorname{rk} W_{k\sim h} = \sum_{x=h+1}^{i} \sum_{y=k}^{l-1} \omega_{yx}$.

Proof. Recall that $\operatorname{rk} W_{k\sim i} = \sum_{x=0}^{i} \sum_{y=k}^{L} \omega_{yx}$. The desired result follows immediately. □

It is also useful to work with the difference of interval diagrams $D = A - B$, where $A \geq_r B$. Denote with $\operatorname{rk} W_{\cdot}$, $\operatorname{rk} V_{\cdot}$, and $\Delta\operatorname{rk} W_{\cdot}$ respectively the ranks of the subsequence matrices of A, B, and D, and denote with $\omega_{\cdot}$, $o_{\cdot}$, and $\Delta\omega_{\cdot}$ the interval multiplicities of A, B, and D respectively. Then, for any $L \geq k \geq i \geq 0$, define $\Delta\omega_{ki} = \omega_{ki} - o_{ki}$, and define $\Delta\operatorname{rk} W_{k\sim i} = \operatorname{rk} W_{k\sim i} - \operatorname{rk} V_{k\sim i}$.

Lemma 30. *Given two interval diagrams A and B such that $A \geq_r B$, the difference interval diagram $D = A - B$ is well-defined. Furthermore, Lemma 29 holds in D.*

Proof. Let $L \geq k \geq i \geq 0$. Since $A \geq_r B$, it is the case that $\Delta\operatorname{rk} W_{k\sim i} \geq 0$ for all choices of k and i by definition. Next, observe that

$$\begin{aligned}
\Delta\operatorname{rk} W_{k\sim i} &= \operatorname{rk} W_{k\sim i} - \operatorname{rk} V_{k\sim i} \\
&= \sum_{x=0}^{i} \sum_{y=k}^{L} \omega_{yx} - \sum_{x=0}^{i} \sum_{y=k}^{L} o_{yx} \\
&= \sum_{x=0}^{i} \sum_{y=k}^{L} (\omega_{yx} - o_{yx}) \\
&= \sum_{x=0}^{i} \sum_{y=k}^{L} \Delta\omega_{yx}.
\end{aligned}$$

This shows that D is a valid interval diagram, and in particular, Lemma 29 holds in it. □

Theorem 31. *Consider two interval diagrams A and B of the same neural network satisfying $A >_r B$, but not necessarily $A >_s B$. We will now show that there exists a sequence of connecting and swapping moves to a reach diagram A' from B such that $A \geq_r A' >_r B$.*

Proof. Let ω_{ki} represent the interval multiplicities for A, and let o_{ki} represent the interval multiplicities for B. Let $L \geq l > k \geq i > h \geq 0$. Let $\Delta\omega_{ki} = \omega_{ki} - o_{ki}$. We will show that if $\Delta\omega_{ki} > 0$ then there exist $l > k$ and $h < i$ such that $\Delta\omega_{lk} < 0$ and $\Delta\omega_{ih} < 0$. This naturally leads to a sequence of swapping and connecting moves that allow us to reach A from B. Let $\Delta\text{rk}\, W_i = \text{rk}\, W_i - \text{rk}\, V_i$, where W_i represents a matrix in A and V_i represents a matrix in B. The property $A >_r B$ is equivalent to the property $\Delta\text{rk}\, W_i \geq 0$ for all $L \geq i \geq 1$.

Let $k \sim i$ be the longest interval with $\Delta\omega_{ki} > 0$ and the smallest k (to break ties). Since $\Delta\omega_{ki} > 0$, it must be the case that $\Delta\text{rk}\, W_{k\sim i} > 0$. If $\Delta\omega_{k-1,i+1} > 0$ then performing the move comb($k \sim i+1, k-1 \sim i$) takes diagram B to A'. Observe that $A' >_r B$ since the rank of $W_{k\sim i}$ increases by exactly 1, and the rank of no other subsequence matrix changes as a result of this move. Furthermore, $A \geq_r A'$ since $\Delta\text{rk}\, W_{k\sim i} \geq 1$. However, if $\Delta\omega_{k-1,i+1} = 0$, then we check if $\Delta\text{rk}\, W_{k\sim i+1} > 0$. If it is the case, then check $\Delta\omega_{k-1,i+2}$. If it is positive then we can perform the move comb($k \sim i+2, k-1 \sim i$). If $\Delta\omega_{k-1,i+2}$ is nonpositive, then we continue with checking $\Delta\text{rk}\, W_{k\sim i+2}$ and so on until $\Delta\text{rk}\, W_{k\sim i+\Delta i} = 0$. Note that this must eventually happen, since $\Delta\text{rk}\, W_{k\sim k} = d_k - d_k = 0$. If $\Delta\omega_{k-1\sim i+\Delta i} > 0$ then we can perform a connecting or swapping move as above. However, when $\Delta\omega_{k-1\sim i+\Delta i} \leq 0$, we know from Lemma 29 that in the difference interval diagram $D = A - B$,

$$\Delta\text{rk}\, W_{k-1\sim i+\Delta i-1} = \Delta\text{rk}\, W_{k\sim i+\Delta i-1} + \Delta\text{rk}\, W_{k-1\sim i+\Delta i} - \Delta\text{rk}\, W_{k\sim i+\Delta i} - \Delta\omega_{k-1,i+\Delta i} \geq 1$$

since $\Delta\text{rk}\, W_{k\sim i+\Delta i-1} \geq 1$ by assumption, $\Delta\text{rk}\, W_{k-1\sim i+\Delta i} \geq 0$ since $A \geq_r B$, $\Delta\text{rk}\, W_{k\sim i+\Delta i} = 0$ by assumption, and $\Delta\omega_{k-1,i+\Delta i} \leq 0$ by assumption. We can show that $\Delta\text{rk}\, W_{k-1\sim i+x-1} \geq 1$ for every x that was traversed in the previous step since since $\Delta\text{rk}\, W_{k\sim i+x-1} \geq 1$ by assumption, $\Delta\text{rk}\, W_{k-1\sim i+\Delta i} \geq 0$ since $A \geq_r B$, $\Delta\text{rk}\, W_{k\sim i+\Delta i} = 0$ by assumption, and $\sum_{z=i+x+1}^{i+\Delta i} \sum_{y=k-1}^{k-1} \Delta\omega_{yx} \leq 0$ since every $\Delta\omega_{yx}$ in the summation is nonpositive by assumption.

Thus we can continue searching for a strictly positive $\Delta\omega$ until we find one and perform a swapping move as above. If we find none and reach a $\Delta\omega_{j-1,j}$ for some j, then we can perform the connecting move comb($k \sim j, j-1 \sim i$). This allows us to reach A' from B. Please refer to Algorithm 2 for a complete description of this algorithm.

Thus whenever we have two interval diagrams $A \geq_r B$, we can reach a diagram A' such that $A' > B$ and $A \geq_r A'$.

□

Lemma 32. *Algorithm 2 runs in time linear in the number of layers in the neural network and always terminates. Furthermore, it returns an interval diagram A' satisfying $A \geq_r A' >_r B$.*

Proof. At each iteration either k' or i' is respectively decremented or incremented, and both variables range between 0 and L, the number of layers. Constant work is done at each iteration. This shows that the number of iterations needed prior to termination is linear in L.

It is clear from Theorem 31 that the interval diagram A' returned by Algorithm 2 satisfies $A \geq_r A' >_r B$. □

Corollary 33. *Whenever we have two diagrams A and B satisfying $A >_r B$, there is a sequence of connecting and swapping moves to reach diagram A from B.*

Proof. One can apply the algorithm in Theorem 31 iteratively until $A =_r A'$ to reach diagram A from diagram B. Due to the strictness of the inequality $A' >_r B$ in the diagram A' reached by the algorithm, each iteration is guaranteed to either make progress until $A =_r A'$. □

Algorithm 2: Interval Shortening

```
begin
    Let k ~ i be the longest interval with Δω_ki > 0 and the smallest k (to break ties)
    Initialize interval diagram A' ← B
    Initialize k' ← k, i' ← i
    Initialize shorten_right ← True
    while k' > i' do
        if Δω_{k'-1,i'+1} > 0 then
            Perform the swapping move between intervals k ~ i' + 1 and k' − 1 ~ i in A'
            return A'
        else if shorten_right and Δrk W_{k'~i'+1} > 0 then
            i' ← i' + 1
        else
            /* We have shortened the longest interval in D = B − A from the
               right side to obtain the leftmost interval for the required
               combinatorial move, now we need to shorten from the left to
               obtain the interval on the right.                              */
            shorten_right ← False

            /* By Theorem 31, we conclude that Δrk W_{k'-1,i} ≥ 1, ..., Δrk W_{k'-1,i'} ≥ 1.
            Hence we can decrement k'.                                        */
            k' ← k' − 1
        end
    end
    if k' = i' then
        Perform the connecting move between intervals k ~ i' and k' ~ i in A'.
    end
    return A'
end
```

A The Dimension of the Nullspace of the Differential

Here we determine the dimension of the nullspace of $\mathrm{d}\mu(\theta)$ (defined in Section 6.4) by relating it to the image of $\mathrm{d}\mu(\theta)$, defined to be

$$\text{image } \mathrm{d}\mu(\theta) = \{\mathrm{d}\mu(\theta)(\Delta\theta) : \Delta\theta \in \mathbb{R}^{d_\theta}\}.$$

This image is always a vector subspace of $\mathbb{R}^{d_h \times d_0}$. The *rank* of image $\mathrm{d}\mu(\theta)$ is the dimension of image $\mathrm{d}\mu(\theta)$. Trager, Kohn, and Bruna [16] show (Lemma 3) that

$$\dim \text{image } \mathrm{d}\mu(\theta) = \sum_{h=1}^{L} \mathrm{rk}\, W_{L\sim h} \cdot \mathrm{rk}\, W_{h-1\sim 0} - \sum_{h=1}^{L-1} \mathrm{rk}\, W_{L\sim h} \cdot \mathrm{rk}\, W_{h\sim 0}.$$

We can determine the dimension of null $\mathrm{d}\mu(\theta)$ from the following observation.

Lemma 34. *The dimensions of the nullspace and image of* $\mathrm{d}\mu(\theta)$ *are related by*

$$\dim \mathrm{null}\, \mathrm{d}\mu(\theta) + \dim \text{image } \mathrm{d}\mu(\theta) = d_\theta.$$

Proof. Let $(\operatorname{null} \mathrm{d}\mu(\theta))^\perp \subseteq \mathbb{R}^{d_\theta}$ denote the subspace of all vectors orthogonal to $\operatorname{null} \mathrm{d}\mu(\theta)$. Then $\dim \operatorname{null} \mathrm{d}\mu(\theta) + \dim(\operatorname{null} \mathrm{d}\mu(\theta))^\perp = d_\theta$. By the following reasoning, $\mathrm{d}\mu(\theta)$ is a bijection from $(\operatorname{null} \mathrm{d}\mu(\theta))^\perp$ to image $\mathrm{d}\mu(\theta)$, so the two subspaces have the same dimension and the result follows.

Every vector $p \in \mathbb{R}^{d_\theta}$ has a unique decomposition $p = p^\parallel + p^\perp$ into a vector $p^\parallel \in \operatorname{null} \mathrm{d}\mu(\theta)$, parallel to the nullspace, and a vector $p^\perp \in (\operatorname{null} \mathrm{d}\mu(\theta))^\perp$, perpendicular to the nullspace. By the linearity of $\mathrm{d}\mu(\theta)$, $\mathrm{d}\mu(\theta)(p) = \mathrm{d}\mu(\theta)(p^\parallel) + \mathrm{d}\mu(\theta)(p^\perp) = \mathrm{d}\mu(\theta)(p^\perp)$. Therefore, $\mathrm{d}\mu(\theta)$ is surjective from $(\operatorname{null} \mathrm{d}\mu(\theta))^\perp$ to image $\mathrm{d}\mu(\theta)$.

For any two distinct points $p_1, p_2 \in (\operatorname{null} \mathrm{d}\mu(\theta))^\perp$, $p_1 - p_2 \in (\operatorname{null} \mathrm{d}\mu(\theta))^\perp \setminus \{0\}$. By the linearity of $\mathrm{d}\mu(\theta)$, $\mathrm{d}\mu(\theta)(p_1) - \mathrm{d}\mu(\theta)(p_2) = \mathrm{d}\mu(\theta)(p_1 - p_2) \neq 0$. Thus $\mathrm{d}\mu(\theta)(p_1) \neq \mathrm{d}\mu(\theta)(p_2)$, and the restriction of $\mathrm{d}\mu(\theta)$ to the domain $(\operatorname{null} \mathrm{d}\mu(\theta))^\perp$ is injective. □

Corollary 35. *The dimension of the nullspace of* $\mathrm{d}\mu(\theta)$ *is*

$$\dim \operatorname{null} \mathrm{d}\mu(\theta) = d_\theta - \sum_{i=1}^{L} \operatorname{rk} W_{L\sim i} \cdot \operatorname{rk} W_{i-1\sim 0} + \sum_{i=1}^{L-1} \operatorname{rk} W_{L\sim i} \cdot \operatorname{rk} W_{i\sim 0}.$$

B Counting More Degrees of Freedom

Table 6 gives the definitions of several prebases that were not important enough to include in Section 6.7, and the dimensions of the subspaces (of $\mathbb{R}^{d_\theta}$) spanned by those prebases.

Let's count connecting moves. Recall from Section 6.3 that a connecting move is a combinatorial move with a sufficiently small displacement $\Delta\theta \in \phi_{lkjih} \setminus \{\mathbf{0}\}$ where $k = j - 1$ and $i = j$. The set of these subspaces is $\Theta_{\mathrm{O}}^{\mathrm{conn}} = \{\phi_{l,j-1,j,j,h} \neq \{\mathbf{0}\} : L \geq l \geq j > h \geq 0\}$.

The total number of degrees of freedom of the infinitesimal connecting moves that change W_j is

$$D_j^{\mathrm{conn}} = \left(\sum_{l=j}^{L} \omega_{lj}\right)\left(\sum_{h=0}^{j-1} \omega_{j-1,h}\right) = \beta_{jjj}\, \alpha_{j-1,j-1,j-1} = (d_j - \operatorname{rk} W_j)\,(d_{j-1} - \operatorname{rk} W_j).$$

Hence, the dimension of the space spanned by $\Theta_{\mathrm{O}}^{\mathrm{conn}}$ is

$$D_{\mathrm{O}}^{\mathrm{conn}} = \sum_{j=1}^{L} D_j^{\mathrm{conn}} = \sum_{j=1}^{L}\left(\sum_{l=j}^{L} \omega_{lj}\right)\left(\sum_{h=0}^{j-1} \omega_{j-1,h}\right) = \sum_{j=1}^{L} \beta_{jjj}\, \alpha_{j-1,j-1,j-1} = \sum_{j=1}^{L} (d_j - \operatorname{rk} W_j)\,(d_{j-1} - \operatorname{rk} W_j).$$

Let's count swapping moves. Recall from Section 6.3 that a swapping move is a combinatorial move with a sufficiently small displacement $\Delta\theta \in \phi_{lkjih} \setminus \{\mathbf{0}\}$ where $k \geq i$ (thereby omitting the connecting moves, which have $k = i - 1$). The set of these subspaces is $\Theta_{\mathrm{O}}^{\mathrm{swap}} = \{\phi_{lkjih} \in \Theta_{\mathrm{O}} : l > k \geq i > h\}$. The dimension $D_{\mathrm{O}}^{\mathrm{swap}}$ of the space spanned by $\Theta_{\mathrm{O}}^{\mathrm{swap}}$ can be derived exactly as we derived $D_{\mathrm{O}}^{\mathrm{comb}}$ in Section 6.7, except that we omit from the count the dimensions of the subspaces where $k = i - 1$. Thus we obtain the formula (22) (see Table 6).

Let's count the degrees of freedom of the infinitesimal combinatorial moves that don't change $\mu(\theta)$. These combinatorial moves move from one stratum to a different stratum of the same fiber. These moves are represented by the prebasis $\Theta_{\mathrm{O}}^{\mathrm{fiber,comb}} = \Theta_{\mathrm{O}}^{\mathrm{fiber}} \cap \Theta_{\mathrm{O}}^{\mathrm{comb}}$. The easiest way to determine the dimension of the subspace spanned by $\Theta_{\mathrm{O}}^{\mathrm{fiber,comb}}$ is to first understand the prebasis $\Theta_{\mathrm{O}}^{L0,\mathrm{comb}} = \Theta_{\mathrm{O}}^{L0} \cap \Theta_{\mathrm{O}}^{\mathrm{comb}}$, which is the set

$$\Theta_{\mathrm{O}}^{\mathrm{conn}} = \{\phi_{l,j-1,j,j,h} \neq \{\mathbf{0}\} : L \geq l \geq j > h \geq 0\}$$

$$D_{\mathrm{O}}^{\mathrm{conn}} = \sum_{L \geq l \geq j > h \geq 0} \omega_{lj}\, \omega_{j-1,h} = \sum_{j=1}^{L} \underbrace{(d_j - \operatorname{rk} W_j)}_{\beta_{jjj}} \underbrace{(d_{j-1} - \operatorname{rk} W_j)}_{\alpha_{j-1,j-1,j-1}}. \tag{21}$$

$$\Theta_{\mathrm{O}}^{\mathrm{swap}} = \{\phi_{lkjih} \in \Theta_{\mathrm{O}} : l > k \geq i > h\} = \{\phi_{lkjih} \neq \{\mathbf{0}\} : L \geq l > k \geq i > h \geq 0 \text{ and } k+1 \geq j \geq i\}$$

$$D_{\mathrm{O}}^{\mathrm{swap}} = \sum_{L > k \geq i > 0} (k - i + 2) \underbrace{(\operatorname{rk} W_{k+1\sim i} - \operatorname{rk} W_{k+1\sim i-1})}_{\beta_{k+1,i,i}} \underbrace{(\operatorname{rk} W_{k\sim i-1} - \operatorname{rk} W_{k+1\sim i-1})}_{\alpha_{k,k,i-1}} \tag{22}$$

$$\Theta_{\mathrm{O}}^{L0,\mathrm{comb}} = \Theta_{\mathrm{O}}^{L0} \cap \Theta_{\mathrm{O}}^{\mathrm{comb}} = \{\phi_{Lkji0} \neq \{\mathbf{0}\} : L \geq k+1 \geq j \geq i > 0\}$$

$$D_{\mathrm{O}}^{L0,\mathrm{comb}} = \sum_{L \geq k+1 \geq i > 0} (k - i + 2)\, \omega_{Li}\, \omega_{k0} \tag{23}$$

$$\Theta_{\mathrm{O}}^{\mathrm{fiber,comb}} = \Theta_{\mathrm{O}}^{\mathrm{fiber}} \cap \Theta_{\mathrm{O}}^{\mathrm{comb}} = \Theta_{\mathrm{O}}^{\mathrm{comb}} \setminus \Theta_{\mathrm{O}}^{L0} = \{\phi_{lkjih} \in \Theta_{\mathrm{O}}^{\mathrm{comb}} : L > l \text{ or } h > 0\}$$

$$D_{\mathrm{O}}^{\mathrm{fiber,comb}} = D_{\mathrm{O}}^{\mathrm{comb}} - D_{\mathrm{O}}^{L0,\mathrm{comb}} = \sum_{L \geq k+1 \geq i > 0} (k - i + 2)\, (\beta_{k+1,i,i}\, \alpha_{k,k,i-1} - \omega_{Li}\, \omega_{k0}) \tag{24}$$

$$\Theta_{\mathrm{O}}^{L0,\mathrm{conn}} = \Theta_{\mathrm{O}}^{L0} \cap \Theta_{\mathrm{O}}^{\mathrm{conn}} = \{\phi_{L,j-1,j,j,0} \neq \{\mathbf{0}\} : L \geq j > 0\}$$

$$D_{\mathrm{O}}^{L0,\mathrm{conn}} = \sum_{j=1}^{L} \omega_{Lj}\, \omega_{j-1,0} = \sum_{j=1}^{L} \underbrace{(\operatorname{rk} W_{L\sim j} - \operatorname{rk} W_{L\sim j-1})}_{\omega_{Lj}} \underbrace{(\operatorname{rk} W_{j-1\sim 0} - \operatorname{rk} W_{j\sim 0})}_{\omega_{j-1,0}} \tag{25}$$

$$\Theta_{\mathrm{O}}^{\mathrm{fiber,conn}} = \Theta_{\mathrm{O}}^{\mathrm{fiber}} \cap \Theta_{\mathrm{O}}^{\mathrm{conn}} = \Theta_{\mathrm{O}}^{\mathrm{conn}} \setminus \Theta_{\mathrm{O}}^{L0} = \{\phi_{l,j-1,j,j,h} \in \Theta_{\mathrm{O}}^{\mathrm{conn}} : L > l \text{ or } h > 0\}$$

$$D_{\mathrm{O}}^{\mathrm{fiber,conn}} = \sum_{j=1}^{L} \left(\underbrace{(d_j - \operatorname{rk} W_j)}_{\beta_{jjj}} \underbrace{(d_{j-1} - \operatorname{rk} W_j)}_{\alpha_{j-1,j-1,j-1}} - \underbrace{(\operatorname{rk} W_{L\sim j} - \operatorname{rk} W_{L\sim j-1})}_{\omega_{Lj}} \underbrace{(\operatorname{rk} W_{j-1\sim 0} - \operatorname{rk} W_{j\sim 0})}_{\omega_{j-1,0}} \right) \tag{26}$$

$$\Theta_{\mathrm{T}}^{L0,\mathrm{comb}} = \Theta_{\mathrm{T}}^{L0} \cap \Theta_{\mathrm{T}}^{\mathrm{comb}} = \{\tau_{Lkji0} \neq \{\mathbf{0}\} : L > k \geq j \geq i > 0\}$$

$$D_{\mathrm{T}}^{L0,\mathrm{comb}} = \sum_{L > k \geq i > 0} (k - i + 1)\, \omega_{Li}\, \omega_{k0} \tag{27}$$

$$\Theta_{\mathrm{T}}^{\mathrm{fiber,comb}} = \Theta_{\mathrm{T}}^{\mathrm{comb}} \setminus \Theta_{\mathrm{T}}^{L0} = \{\tau_{lkjih} \in \Theta_{\mathrm{T}}^{\mathrm{comb}} : L > l \text{ or } h > 0\}$$

$$D_{\mathrm{T}}^{\mathrm{fiber,comb}} = D_{\mathrm{T}}^{\mathrm{comb}} - D_{\mathrm{T}}^{L0,\mathrm{comb}} = \sum_{L > k \geq i > 0} (k - i + 1)\, (\beta_{k+1,i,i}\, \alpha_{k,k,i-1} - \omega_{Li}\, \omega_{k0}) \tag{28}$$

Table 6: More sets of subspaces of moves and their total degrees of freedom. See also Table 5.

of one-matrix subspaces representing combinatorial moves that change $\mu(\theta)$ (don't stay on the fiber). As $\Theta_{\mathrm{O}}^{L0,\mathrm{comb}} = \{\phi_{Lkji0} \neq \{\mathbf{0}\} : L \geq k+1 \geq j \geq i > 0\}$, it spans a subspace of dimension

$$D_{\mathrm{O}}^{L0,\mathrm{comb}} = \sum_{L \geq k+1 \geq j \geq i > 0} \omega_{Li}\, \omega_{k0} = \sum_{L \geq k+1 \geq i > 0} (k - i + 2)\, \omega_{Li}\, \omega_{k0},$$

because the term $\omega_{Li}\, \omega_{k0}$ appears in the first summation once for each $j \in [i, k+1]$. Analogously, for the set $\Theta_{\mathrm{T}}^{L0,\mathrm{comb}} = \Theta_{\mathrm{T}}^{L0} \cap \Theta_{\mathrm{T}}^{\mathrm{comb}}$ of two-matrix subspaces,

$$D_{\mathrm{T}}^{L0,\mathrm{comb}} = \sum_{L > k \geq j \geq i > 0} \omega_{Li}\, \omega_{k0} = \sum_{L > k \geq i > 0} (k - i + 1)\, \omega_{Li}\, \omega_{k0}.$$

The dimension of the space spanned by $\Theta_{\mathrm{O}}^{\mathrm{fiber,comb}}$ is

$$\begin{aligned} D_{\mathrm{O}}^{\mathrm{fiber,comb}} &= \dim \Theta_{\mathrm{O}}^{\mathrm{fiber,comb}} \\ &= \dim \left(\Theta_{\mathrm{O}}^{\mathrm{comb}} \setminus \Theta_{\mathrm{O}}^{L0}\right) \\ &= \dim \left(\Theta_{\mathrm{O}}^{\mathrm{comb}} \setminus \Theta_{\mathrm{O}}^{L0,\mathrm{comb}}\right) \\ &= D_{\mathrm{O}}^{\mathrm{comb}} - D_{\mathrm{O}}^{L0,\mathrm{comb}}, \end{aligned}$$

from which we obtain the formula (24) in Table 6. Analogously, we define $\Theta_{\mathrm{T}}^{\mathrm{fiber,comb}} = \Theta_{\mathrm{T}}^{\mathrm{comb}} \setminus \Theta_{\mathrm{T}}^{L0}$, which spans a space of dimension $D_{\mathrm{T}}^{\mathrm{fiber,comb}} = D_{\mathrm{T}}^{\mathrm{comb}} - D_{\mathrm{T}}^{L0,\mathrm{comb}}$, from which we obtain the formula (28).

A connecting move that changes W_j also changes $\mu(\theta)$ if and only if $l = L$ and $h = 0$; that is, $\Delta\theta \in \phi_{L,j-1,j,j,0} \setminus \{\mathbf{0}\}$. The total degrees of freedom of the connecting moves that change both W_j and $\mu(\theta)$ is

$$D_j^{L0,\mathrm{conn}} = \omega_{Lj}\,\omega_{j-1,0} = (\mathrm{rk}\, W_{L\sim j} - \mathrm{rk}\, W_{L\sim j-1})\,(\mathrm{rk}\, W_{j-1\sim 0} - \mathrm{rk}\, W_{j\sim 0}).$$

The dimension of the space spanned by $\Theta_{\mathrm{O}}^{L0,\mathrm{conn}} = \Theta_{\mathrm{O}}^{L0} \cap \Theta_{\mathrm{O}}^{\mathrm{conn}}$ is

$$D_{\mathrm{O}}^{L0,\mathrm{conn}} = \sum_{j=1}^{L} D_j^{L0,\mathrm{conn}} = \sum_{j=1}^{L} \omega_{Lj}\,\omega_{j-1,0} = \sum_{j=1}^{L} (\mathrm{rk}\, W_{L\sim j} - \mathrm{rk}\, W_{L\sim j-1})\,(\mathrm{rk}\, W_{j-1\sim 0} - \mathrm{rk}\, W_{j\sim 0}).$$

Let $\Theta_{\mathrm{O}}^{\mathrm{fiber,conn}} = \Theta_{\mathrm{O}}^{\mathrm{fiber}} \cap \Theta_{\mathrm{O}}^{\mathrm{conn}} = \Theta_{\mathrm{O}}^{\mathrm{conn}} \setminus \Theta_{\mathrm{O}}^{L0}$ represent the connecting moves that stay on the fiber. The dimension of the space spanned by $\Theta_{\mathrm{O}}^{\mathrm{fiber,conn}}$ is

$$\begin{aligned} D_{\mathrm{O}}^{\mathrm{fiber,conn}} &= D_{\mathrm{O}}^{\mathrm{conn}} - D_{\mathrm{O}}^{L0,\mathrm{conn}} \\ &= \sum_{j=1}^{L} \Big((d_j - \mathrm{rk}\, W_j)\,(d_{j-1} - \mathrm{rk}\, W_j) - (\mathrm{rk}\, W_{L\sim j} - \mathrm{rk}\, W_{L\sim j-1})\,(\mathrm{rk}\, W_{j-1\sim 0} - \mathrm{rk}\, W_{j\sim 0}) \Big). \end{aligned}$$

Perhaps it is worth noting that the triple summation in (12) can be simplified to a double summation, because the term $\omega_{Li}\,\omega_{k0}$ occurs once for each $j \in [\max\{i,1\}, \min\{k+1, L\}]$. In (16), the term $\omega_{Li}\,\omega_{k0}$ occurs once for each $j \in [\max\{i,1\}, \min\{k, L-1\}]$. Hence, we can write

$$\begin{aligned} D_{\mathrm{O}}^{L0} &= \sum_{i=0}^{L} \sum_{k=\max\{i-1,0\}}^{L} (\min\{k+1, L\} - \max\{i-1, 0\})\,\omega_{Li}\,\omega_{k0} \quad \text{and} \\ D_{T}^{L0} &= \sum_{i=0}^{L-1} \sum_{k=\max\{i,1\}}^{L} (\min\{k, L-1\} - \max\{i-1, 0\})\,\omega_{Li}\,\omega_{k0}. \end{aligned}$$

References

[1] Sanjeev Arora, Nadav Cohen, and Elad Hazan. *On the optimization of deep networks: Implicit acceleration by overparameterization*. CoRR **abs/1802.06509**, 2018.

[2] Sanjeev Arora, Nadav Cohen, Wei Hu, and Yuping Luo. *Implicit regularization in deep matrix factorization*. CoRR **abs/1905.13655**, 2019.

[3] Anna Choromanska, Mikael Henaff, Michaël Mathieu, Gérard Ben Arous, and Yann LeCun. *The loss surface of multilayer networks*. CoRR **abs/1412.0233**, 2014.

[4] Hung-Hsu Chou, Carsten Gieshoff, Johannes Maly, and Holger Rauhut. *Gradient descent for deep matrix factorization: Dynamics and implicit bias towards low rank*. CoRR **abs/2011.13772**, 2020.

[5] Georg Frobenius. *Sitzungsberichte der Königlich Preussischen*. Akademie de Wissenschaften zu Berlin:20–29, 128–129, 1911. Reprinted in *Ferdinand Georg Frobenius Gesammelte Abhandlungen* **III**:479–490, Springer-Verlag (Berlin), 1968.

[6] Suriya Gunasekar, Blake Woodworth, Srinadh Bhojanapalli, Behnam Neyshabur, and Nathan Srebro. *Implicit regularization in matrix factorization*, 2017.

[7] Moritz Hardt and Tengyu Ma. *Identity matters in deep learning*. CoRR **abs/1611.04231**, 2016.

[8] Ziwei Ji and Matus Telgarsky. *Gradient descent aligns the layers of deep linear networks*. CoRR **abs/1810.02032**, 2018.

[9] Kenji Kawaguchi. *Deep learning without poor local minima*, 2016.

[10] Qianyi Li and Haim Sompolinsky. *Statistical mechanics of deep linear neural networks: The back-propagating renormalization group*. CoRR **abs/2012.04030**, 2020.

[11] Haihao Lu and Kenji Kawaguchi. *Depth creates no bad local minima*. CoRR **abs/1702.08580**, 2017.

[12] Adityanarayanan Radhakrishnan, Eshaan Nichani, Daniel Irving Bernstein, and Caroline Uhler. *Balancedness and alignment are unlikely in linear neural networks*. CoRR **abs/2003.06340**, 2020.

[13] Andrew Saxe, James McClelland, and Surya Ganguli. *Exact solutions to the nonlinear dynamics of learning in deep linear neural networks*. 12 2013.

[14] Gilbert Strang. *The Fundamental Theorem of Linear Algebra*. The American Mathematical Monthly **100**(9):848–855, November 1993.

[15] Néstor Thome. *Inequalities and Equalities for $\ell = 2$ (Sylvester), $\ell = 3$ (Frobenius), and $\ell > 3$ Matrices*. Aequationes Mathematicae **90**(5):951–960, 2016.

[16] Matthew Trager, Kathlén Kohn, and Joan Bruna. *Pure and Spurious Critical Points: A Geometric Study of Linear Networks*. Eighth International Conference on Learning Representations (Addis Ababa, Ethiopia), April 2020.

[17] Li Zhang. *Depth creates no more spurious local minima*. CoRR **abs/1901.09827**, 2019.