The Geometry of the Set of Equivalent Linear Neural Networks

Jonathan Richard Shewchuk and Sagnik Bhattacharya

April 24, 2024

Department of Electrical Engineering and Computer Sciences
University of California at Berkeley
Berkeley, California 94720

Abstract

We characterize the geometry and topology of the set of all weight vectors for which a linear neural network computes the same linear transformation W. This set of weight vectors is called the *fiber* of W (under the matrix multiplication map), and it is embedded in the Euclidean *weight space* of all possible weight vectors. The fiber is an algebraic variety that is not necessarily a manifold. We describe a natural way to *stratify* the fiber—that is, to partition the algebraic variety into a finite set of manifolds of varying dimensions called *strata*. We call this set of strata the *rank stratification*. We derive the dimensions of these strata and the relationships by which they adjoin each other. Although the strata are disjoint, their closures are not. Our strata satisfy the *frontier condition*: if a stratum intersects the closure of another stratum, then the former stratum is a subset of the closure of the latter stratum. Each stratum is a manifold of class C^∞ embedded in weight space, so it has a well-defined tangent space and normal space at every point (weight vector). We show how to determine the subspaces tangent to and normal to a specified stratum at a specified point on the stratum, and we construct elegant bases for those subspaces.

We define transformations in weight space called *moves*, which map one weight vector to another on the same fiber, thereby modifying the neural network's weights without changing the linear transformation that the network computes. Some moves stay on the same stratum. Some moves move from one stratum to another stratum of the fiber; these moves give us useful intuition about how strata adjoin each other. Moves also have a practical use: we can visit different weight vectors for which the neural network computes the same transformation. Some of these weight vectors are better behaved than others in gradient descent algorithms.

To help achieve these goals, we first derive what we call a *Fundamental Theorem of Linear Neural Networks*, analogous to what Strang calls the *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. The topology of a stratum depends solely on its basis flow diagram. So does its geometry, up to a linear transformation in weight space.

Contents

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. We write the composition as

$$W = \mu(W_L, W_{L-1}, \ldots, W_2, W_1) = W_L W_{L-1} \cdots W_2 W_1,$$

where μ is called the *matrix multiplication map*. The network takes an *input vector x* and produces an *output vector* $y = W_L W_{L-1} \cdots W_2 W_1 x$. We number the matrices in the order they are applied in computation. Each matrix W_j is interpreted as a layer of edges (connections) in the network, edge layer number j, and each component of W_j is interpreted as the *weight* of an edge. For brevity, we omit added terms, which do not appreciably affect our results.

This definition is incomplete: a *neural network* also entails software that computes the sequence of vectors, an optimization algorithm that *trains* the network by choosing good weights, and more. To a computer scientist, 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.)

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 a *real algebraic variety*—the set of all real-valued solutions of a system of polynomial equations. The set $\mu^{-1}(W)$ is called the *fiber* of W under the map μ, and we will simply call it the fiber, adopting a convention of Trager, Kohn, and Bruna [21], whose paper inspired this one. Said differently, we 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: it is a union of manifolds of various 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.

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 composed of 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

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

This fiber is a smooth, $(d^2(L-1))$-dimensional manifold with topology $\mathrm{GL}(d, \mathbb{R}) \times \mathrm{GL}(d, \mathbb{R}) \times \ldots \times \mathrm{GL}(d, \mathbb{R})$ (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.

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 an example, Figure 2 graphs the solutions of $[\theta_2][\theta_1 \quad \theta_1'] = [0 \quad 0]$ (an instantiation of $W_2 W_1 = W$). The set of solutions can be 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 θ_1-θ_1' plane with the origin removed. The two subscripts of S_{**} are the ranks of W_2 and W_1, respectively. Clearly, the strata in one fiber can have different dimensions. The stratum S_{00} lies in the closures of both S_{10} and S_{01}, and we think of S_{00} as a branching point connecting S_{10} to S_{01}.
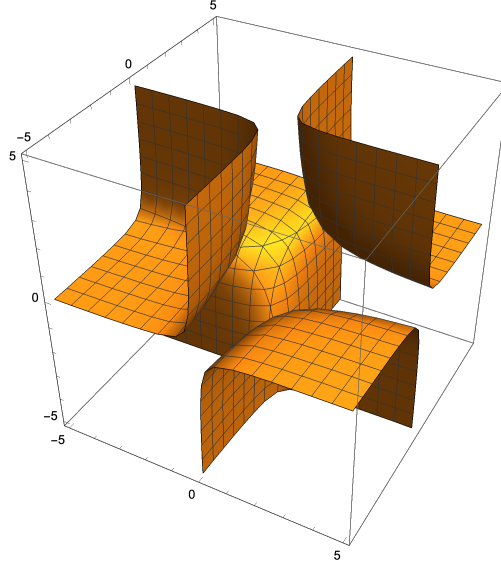
Figure 1: The fiber $\mu^{-1}([1])$ for the network $W_3 W_2 W_1 = [\theta_3][\theta_2][\theta_1] = [1] = W$.
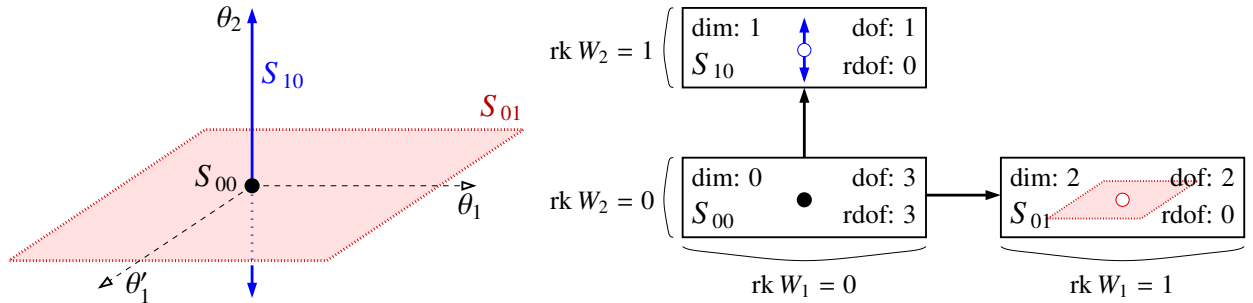
Figure 2: At left is the fiber $\mu^{-1}([0\ \ 0])$ for the network $W_2 W_1 = [\theta_2][\theta_1\ \ \theta_1'] = [0\ \ 0] = W$, partitioned into three strata: S_{00} is the origin; S_{10} (blue) is the θ_2-axis with the origin removed; and S_{01} (pink) is the plane spanned by the θ_1- and θ_1'-axes with the origin removed. At right, the strata are arranged in a *stratum dag*, which we organize 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), which stay on the fiber but move off the stratum and onto a higher-dimensional stratum. Always, dof = dim + rdof. A directed path from one stratum to another implies that the former is a subset of the closure of the latter.

Each stratum represents a different pattern by which information flows (or fails to flow) through the neural network. To understand information flow in a linear neural network, we derive a *Fundamental Theorem of Linear Neural Networks*, analogous to what Strang [18] calls the *Fundamental Theorem of Linear Algebra*. If we think of a matrix as a linear map, its domain can be decomposed into its rowspace and its nullspace, and its range can be decomposed into its columnspace and its left nullspace. In a linear neural network, there is an analogous decomposition of each layer of units into a set of subspaces, determined by observing which information is annihilated at which layers of the network. The information carried in a particular subspace is *all* lost in the nullspace of the *same* layer of edges, *or* it all reaches the network's output. Symmetrically, either the subspace carries information from the input or it merely has the potential to carry information from some earlier layer—a potential that may be realized 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. The topology of a stratum depends solely on the basis flow diagram associated with the stratum. So does its geometry, up to a linear transformation in weight space.

The bulk of this paper is spent explaining the geometry of the strata and how the strata are connected to each other. We will determine the tangent spaces and normal spaces of each stratum by using the subspaces in the basis flow diagrams as building blocks. We also study a set of operations we call *moves* that map one network factorization of W to another—in effect, moving along the fiber. This study has both theoretical and practical motivations. The theoretical motivation is that moves provide a great deal of intuition about the geometry and topology of the fiber of a matrix W. In particular, some moves reveal how the strata are connected to each other. 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's weight vector 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 the network's layers [21]. Spurious critical points appear to be one of the reasons that deep neural networks sometimes learn more slowly than shallow ones. Our original motivation for this paper is that we want to find practical 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 is a 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 [1, 2, 4, 5, 8, 10–12, 14, 17, 21, 25]. Why? We cannot fully understand the training of ReLU-based networks—or probably any neural networks—if we do not understand linear networks. Researchers have studied linear neural networks to improve our understanding of observed phenomena in ReLU networks such as implicit regularization in minimizing the training algorithm's cost function [2, 5, 7], implicit acceleration of training by gradient descent [1], and the success of residual networks [8]. Close to our hearts, Trager, Kohn, and Bruna [21] 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.

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 $\mathbb{R}^{d_j}$. Unit layer 0 is the *input layer*, unit 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 matrix multiplication map $\mu(W_L, W_{L-1}, \ldots, W_2, W_1) = W_L W_{L-1} \cdots W_2 W_1$; we can abbreviate it to $\mu(\theta)$. Given a fixed weight vector θ, the 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 (unless $L = 1$), so we define its preimage to be a set. Given $W \in \mathbb{R}^{d_L \times d_0}$, 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 under μ, and we treat it as a geometric object in $\mathbb{R}^{d_\theta}$. Note that $\mu^{-1}(W)$ is empty if and only if rk $W > \min_{j=1}^{L-1} d_j$.

Given $\theta = (W_L, W_{L-1}, \ldots, W_1) \in \mathbb{R}^{d_\theta}$, its *subsequence matrices* are all the matrices of the form

$$W_{k \sim i} = \mu_{k \sim i}(\theta), \quad \text{where} \quad \mu_{k \sim i}(\theta) = W_k W_{k-1} \cdots W_{i+1}, \quad L \ge k \ge i \ge 0.$$

The notation $W_{k \sim i}$ (and $\mu_{k \sim i}$) 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 use the convention that $W_{k \sim k} = I_{d_k \times d_k}$, the $d_k \times d_k$ identity matrix. Throughout this paper, assume that every $W_{k \sim i}$ is a function of θ unless otherwise stated. We call each W_j a *factor matrix*.

The *rank list* $\underline{r}$ 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 $L \ge k \ge i \ge 0$. The list includes the unit layer sizes rk $W_{k \sim k} = d_k$. For example, for a network with $L = 3$ layers of edges, the rank list is

$$\underline{r} = \langle d_3, d_2, d_1, d_0, \text{rk } W_3, \text{rk } W_2, \text{rk } W_1, \text{rk } W_3 W_2, \text{rk } W_2 W_1, \text{rk } W \rangle.$$

We will use rank lists to partition a fiber into strata, which we will show are smooth manifolds. Sometimes we do not want to specify a particular θ, but rather we wish to specify some target ranks. In that case, we let $r_{k \sim i}$ denote the target value of rk $W_{k \sim i}$ and we write $\underline{r} = \langle r_{k \sim i} \rangle_{L \ge k \ge i \ge 0}$. For example, if we set $r_{4 \sim 2} = 2$, we select weight vectors for which rk $W_4 W_3 = 2$.

Let $S_{\underline{r}}^W$ denote the set of points in W's fiber having rank list $\underline{r}$. That is,

$$S_{\underline{r}}^W = \{\theta \in \mu^{-1}(W) : \text{the rank list of } \theta \text{ is } \underline{r}\}.$$

We call $S_{\underline{r}}^W$ a *stratum* in the *rank stratification* of W's fiber. When W is clear from context, we just write $S_{\underline{r}}$.

Given two rank lists $\underline{r}$ and $\underline{s}$, we write $\underline{r} \le \underline{s}$ to mean that $r_{k \sim i} \le s_{k \sim i}$ for all $L \ge k \ge i \ge 0$. We write $\underline{r} < \underline{s}$ to mean that $\underline{r} \le \underline{s}$ and $\underline{r} \ne \underline{s}$ (at least one of the inequalities holds strictly). For example, $\langle 5, 4, 1 \rangle < \langle 5, 4, 3 \rangle$. Given a set $S \subseteq \mathbb{R}^{d_\theta}$, let $\bar{S}$ denote the closure of S (with respect to the weight space $\mathbb{R}^{d_\theta}$). We will see that for a nonempty stratum $S_{\underline{r}}$, $S_{\underline{r}} \subseteq \bar{S}_{\underline{s}}$ if and only if $\underline{r} \le \underline{s}$.

Given a point (weight vector) θ on a smooth manifold $S \subseteq \mathbb{R}^{d_\theta}$, $T_\theta S$ denotes the *tangent space* of S at θ—the subspace of $\mathbb{R}^{d_\theta}$ tangent to S at θ that has the same dimension as S—and $N_\theta S$ denotes its orthogonal complement, the *normal space* of S at θ—the subspace of $\mathbb{R}^{d_\theta}$ orthogonal to S at θ that has dimension $d_\theta - \dim S$. In our setting, both $T_\theta S$ and $N_\theta S$ pass through the origin (they are true subspaces of $\mathbb{R}^{d_\theta}$), not necessarily through θ. Note that two vectors $\phi = (X_L, X_{L-1}, \ldots, X_1)$ and $\psi = (Y_L, Y_{L-1}, \ldots, Y_1)$ in the weight space $\mathbb{R}^{d_\theta}$ are *orthogonal* if $\phi \cdot \psi = 0$; hence $\phi \cdot \psi = 0$ for every $\phi \in T_\theta S$ and $\psi \in N_\theta S$. This Euclidean inner product can be written $\phi \cdot \psi = \sum_{j=1}^L \langle X_j, Y_j \rangle_F$, where $\langle X, Y \rangle_F = \sum_{i=1}^p \sum_{k=1}^q X_{ik} Y_{ik}$ denotes the Frobenius inner product of two $p \times q$ matrices X and Y.

We have a frequent need to multiply a matrix by a subspace. Given a $p \times q$ matrix M and a subspace $Z \in \mathbb{R}^q$, we define

$$MZ = \{Mv : v \in Z\},$$

which is a subspace of $\mathbb{R}^p$. For example, $M\mathbb{R}^q$ is the columnspace of M.

Tables 1 and 2 collect most of the notation used in this paper for easy reference.

A_{kji} = null $W_{k+1 \sim j} \cap$ col $W_{j \sim i}$ flow subspace of $\mathbb{R}^{d_j}$ for unit layer j (note: $A_{j-1,j,i} = \{\mathbf{0}\} = A_{k,j,-1}$)

$a_{kji} \in A_{kji} \downarrow (A_{k,j,i-1} + A_{k-1,j,i})$ prebasis subspace of $\mathbb{R}^{d_j}$ for unit layer j

$\mathcal{A}_{kji} = \{a_{k'ji'} \neq \mathbf{0}\} : k' \in [j,k], i' \in [0,i]\}$ prebasis for A_{kji} (set of prebasis subspaces spanning A_{kji})

$\mathcal{A}_j = \mathcal{A}_{Ljj}$ prebasis for $\mathbb{R}^{d_j}$ for unit layer j

B_{kji} = row $W_{k \sim j} \cap$ null $W_{j \sim i-1}^{\mathsf{T}}$ flow subspace of $\mathbb{R}^{d_j}$ for unit layer j of the transpose network

$b_{kji} \in B_{kji} \downarrow (B_{k,j,i+1} + B_{k+1,j,i})$ prebasis subspace of $\mathbb{R}^{d_j}$ for unit layer j of the transpose network

$\mathcal{B}_{kji} = \{b_{k'ji'} \neq \mathbf{0}\} : k' \in [k,L], i' \in [i,j]\}$ prebasis for B_{kji} (set of prebasis subspaces spanning B_{kji})

$\mathcal{B}_j = \mathcal{B}_{jj0}$ prebasis for $\mathbb{R}^{d_j}$ for unit layer j of the transpose network

d_j number of units in unit layer j of the neural network

$d_\theta = d_L d_{L-1} + d_{L-1} d_{L-2} + \ldots + d_1 d_0$ number of weights in network (dimension of the weight space)

$D_{\mathrm{O}}^{L0}, D_{\mathrm{O}}^{\mathrm{fiber}}, D_{\mathrm{O}}^{\mathrm{comb}}, D_{\mathrm{O}}^{\mathrm{stratum}}, D_{\mathrm{T}}^{L0}, D_{\mathrm{T}}^{\mathrm{comb}}, D^{\mathrm{free}}, D^{\mathrm{stratum}}, D^{\mathrm{fiber}}$ dimension of span Θ_{O}^{L0}, etc.; Table 6

$e_{lkyxih} = b_{lyi} \otimes a_{kxh}$ subsequence subspace of $\mathbb{R}^{d_y \times d_x}$

$\mathcal{E}_{yx} = \{e_{lkyxih} \neq \mathbf{0}\} : l \in [y,L], k \in [x,L], i \in [0,y], h \in [0,x]\}$ subsequence prebasis for $\mathbb{R}^{d_y \times d_x}$

$\tilde{I}_j$ almost-identity matrix (corresponding to the factor matrix W_j)

J_{kji} a $d_j \times \omega_{ki}$ matrix whose columns are a basis for a_{kji}

$J_j = [J_{kji}]_{k \in [j,L], i \in [0,j]}$ invertible $d_j \times d_j$ matrix whose columns are a basis for $\mathbb{R}^{d_j}$ for unit layer j

K_{kji} a $d_j \times \omega_{ki}$ matrix whose columns are a basis for b_{kji}

$K_j = [K_{kji}]_{k \in [j,L], i \in [0,j]}$ invertible $d_j \times d_j$ matrix whose columns are a basis for $\mathbb{R}^{d_j}$ for unit layer j

L number of layers of edges in the neural network

$N_j =$ null $W_{L \sim j} \otimes \mathbb{R}^{d_{j-1}} + \mathbb{R}^{d_j} \otimes$ null $W_{j-1 \sim 0}^{\mathsf{T}}$ set of displacements of W_j that do not change W

$N_\theta S$ subspace of $\mathbb{R}^{d_\theta}$ normal to manifold S at θ (passes through origin, not necessarily through θ)

$o_{lkjih} = a_{lji} \otimes b_{k,j-1,h}$ prebasis subspace of $\mathbb{R}^{d_j \times d_{j-1}}$

$O_j = \{o_{lkjih} \neq \mathbf{0}\} : l \in [j,L], k \in [j-1,L], i \in [0,j], h \in [0,j-1]\}$ prebasis for $\mathbb{R}^{d_j \times d_{j-1}}$

$O_j^{L0} = \{o_{lkjih} \in O_j : l = L \text{ and } h = 0\}$ subspaces in O_j that are linearly independent of N_j

$O_j^{\mathrm{fiber}} = O_j \setminus O_j^{L0}$ subspaces in O_j that are subsets of N_j

$P = \{(W_L, W_{L-1}, \ldots, W_{j+2}, W_{j+1}(I + \epsilon H), (I + \epsilon H)^{-1} W_j, W_{j-1}, \ldots, W_1) : \epsilon \in [0, \hat{\epsilon}]\}$ two-matrix path

$\underline{r} = \langle \mathrm{rk}\, W_{k \sim i} \rangle_{L \geq k \geq i \geq 0}$ or $\langle r_{k \sim i} \rangle_{L \geq k \geq i \geq 0}$ rank list; lists the ranks of all subsequence matrices $W_{k \sim i}$

$r_{k \sim i}$ target value for rk $W_{k \sim i}$; one of the entries in rank list $\underline{r}$

$S_{\underline{r}} = S_{\underline{r}}^W = \{\theta \in \mu^{-1}(W) : \text{the rank list of } \theta \text{ is } \underline{r}\}$ stratum (of the fiber of W) with rank list $\underline{r}$

$\mathcal{S} = \{S_{\underline{r}} : \underline{r} \text{ is the rank list of some weight vector in } \mu^{-1}(W)\}$ rank stratification of $\mu^{-1}(W)$

$T_\theta S$ subspace of $\mathbb{R}^{d_\theta}$ tangent to manifold S at θ (passes through origin, not necessarily through θ)

$W = \mu(\theta) = W_L W_{L-1} \cdots W_1$ product of factor matrices; linear map computed by neural network

$W_j \in \mathbb{R}^{d_j \times d_{j-1}}$ factor matrix; layer j of edges in neural network; connects unit layers $j-1$ and j

ΔW_j displacement applied to factor matrix W_j

$W_j' = W_j + \Delta W_j$ factor matrix after a move from $\theta = (W_L, \ldots, W_1)$ to $\theta' = (W_L', \ldots, W_1')$

$W_{k \sim i} = \mu_{k \sim i}(\theta) = W_k W_{k-1} \cdots W_{i+1}$ subsequence matrix; maps unit layer i to unit layer k

     Notes: by convention, $W_{k \sim k} = I_{d_k \times d_k}$; $W_{k \sim -1} = 0_{d_k \times 1}$; $W_{L+1 \sim i} = 0_{1 \times d_i}$; $W_{L \sim 0} = W$; $W_{j \sim j-1} = W_j$

$x \in \mathbb{R}^{d_0}$ input vector

$y \in \mathbb{R}^{d_L}$ output vector; $y = Wx = W_L W_{L-1} \cdots W_1 x$

$\mathrm{DV}_r^{p \times q} = \{M \in \mathbb{R}^{p \times q} : \mathrm{rk}\, M \leq r\}$ determinantal variety

$\mathrm{DM}_r^{p \times q} = \{M \in \mathbb{R}^{p \times q} : \mathrm{rk}\, M = r\} = \mathrm{DV}_r^{p \times q} \setminus \mathrm{DV}_{r-1}^{p \times q}$ determinantal manifold

$\mathrm{WDM}_{\underline{r}}^{k \sim i} = \{(W_L, \ldots, W_1) \in \mathbb{R}^{d_\theta} : \mathrm{rk}\, W_{k \sim i} = r_{k \sim i}\}$ weight-space determinantal manifold

$\mathrm{MDM}_{\underline{r}} = \{(W_L, \ldots, W_1) \in \mathbb{R}^{d_\theta} : \mathrm{rk}\, W_{k \sim i} = r_{k \sim i} \text{ for all } L \geq k \geq i \geq 0\}$ multideterminantal manifold

     Notes: $\mathrm{MDM}_{\underline{r}} = \bigcap_{L \geq k > i \geq 0} \mathrm{WDM}_{\underline{r}}^{k \sim i}$ and $S_{\underline{r}}^W = \mu^{-1}(W) \cap \mathrm{MDM}_{\underline{r}}$

$\mathrm{GL}(d, \mathbb{R})$ the general linear group on real-valued $d \times d$ matrices

Table 1: Notation used in this paper. Continued in Table 2. See also Table 6.

$\alpha_{kji} = \dim A_{kji} = \sum_{t=j}^{k} \sum_{s=0}^{i} \omega_{ts}$ dimension of flow subspace A_{kji}

$\beta_{kji} = \dim B_{kji} = \sum_{t=k}^{L} \sum_{s=i}^{j} \omega_{ts}$ dimension of flow subspace B_{kji}

$\zeta_j = \{(W_L, W_{L-1}, \dots, W_{j+1}, W_j + \Delta W_j, W_{j-1}, \dots, W_1) : \Delta W_j \in N_j\}$ weight-space version of N_j

$\eta(M_L, M_{L-1}, \dots, M_1) = (J_L M_L J_{L-1}^{-1}, J_{L-1} M_{L-1} J_{L-2}^{-1}, \dots, J_1 M_1 J_0^{-1})$ a linear transformation of $\mathbb{R}^{d_\theta}$

$\theta = (W_L, W_{L-1}, \dots, W_1) \in \mathbb{R}^{d_\theta}$ weight vector representing a neural network

$\Delta\theta = (\Delta W_L, \Delta W_{L-1}, \dots, \Delta W_1) \in \mathbb{R}^{d_\theta}$ displacement applied to a weight vector θ

$\theta' = \theta + \Delta\theta$ weight vector after a move from θ to θ'

$\tilde{\theta} = (\tilde{I}_L, \tilde{I}_{L-1}, \dots, \tilde{I}_1)$ the canonical weight vector (for some specified rank list $\underline{r}$)

$\Theta_O = \{\phi_{lkjih} \neq \{\mathbf{0}\} : L \geq l \geq j \geq i \geq 0 \text{ and } L \geq k \geq j - 1 \geq h \geq 0\}$ the one-matrix prebasis for $\mathbb{R}^{d_\theta}$

$\Theta_T = \{\tau_{lkjih} \neq \{\mathbf{0}\} : L \geq l > j > h \geq 0 \text{ and } L \geq k \geq j \geq i \geq 0\}$ restricted set of two-matrix subspaces

$\Theta_{T+} = \{\tau_{lkjih} \neq \{\mathbf{0}\} : L > j > 0, L \geq l \geq j \geq h \geq 0, \text{ and } L \geq k \geq j \geq i \geq 0\}$ two-matrix subspaces

$\Theta_O^{L0}, \Theta_O^{\text{fiber}}, \Theta_O^{\text{comb}}, \Theta_O^{\text{stratum}}, \Theta_T^{L0}, \Theta_T^{\text{comb}}, \Theta^{\text{free}}, \Theta^{\text{stratum}}, \Theta^{\text{fiber}}$ prebases for subspaces of $\mathbb{R}^{d_\theta}$; Table 6

$\mu(\theta) = \mu(W_L, W_{L-1}, \dots, W_1) = W_L W_{L-1} \cdots W_1$ matrix multiplication map, $\mu : \mathbb{R}^{d_\theta} \to \mathbb{R}^{d_L \times d_0}$

$\mu^{-1}(W) = \{\theta \in \mathbb{R}^{d_\theta} : \mu(\theta) = W\}$ the fiber of W under the matrix multiplication map, $W \in \mathbb{R}^{d_L \times d_0}$

$d\mu(\theta)(\Delta\theta) = \sum_{j=1}^{L} W_{L \sim j} \Delta W_j W_{j-1 \sim 0}$ the differential map of μ at θ is $d\mu(\theta) : \mathbb{R}^{d_\theta} \to \mathbb{R}^{d_L \times d_0}$

$d\mu^\top(\theta)(\Delta W) = (\Delta W W_{L-1 \sim 0}^\top, \dots, W_{L \sim j}^\top \Delta W W_{j-1 \sim 0}^\top, \dots, W_{L \sim 1}^\top \Delta W)$ transpose of the differential map

$\mu_{y \sim x}(\theta) = W_y W_{y-1} \cdots W_{x+1}$ subsequence matrix $W_{y \sim x}$'s map, $\mu_{y \sim x} : \mathbb{R}^{d_\theta} \to \mathbb{R}^{d_y \times d_x}$

$d\mu_{y \sim x}(\theta)(\Delta\theta) = \sum_{j=x+1}^{y} W_{y \sim j} \Delta W_j W_{j-1 \sim x}$ the differential map of $\mu_{y \sim x}$ at θ is $d\mu_{y \sim x}(\theta) : \mathbb{R}^{d_\theta} \to \mathbb{R}^{d_y \times d_x}$

$\tau_{lkjih} = \{(0, 0, \dots, 0, W_{j+1}H, -HW_j, 0, \dots, 0) : H \in a_{lji} \otimes b_{kjh}\}$ two-matrix subspace of $\mathbb{R}^{d_\theta}$

$\phi_{lkjih} = \{(0, 0, \dots, 0, \Delta W_j, 0, \dots, 0) : \Delta W_j \in o_{lkjih} = a_{lji} \otimes b_{k,j-1,h}\}$ one-matrix subspace of $\mathbb{R}^{d_\theta}$

$\chi_r(M)$ maps matrix M to a vector listing the determinant of every $(r+1) \times (r+1)$ minor of M

$\psi_{lkyxih} = \{(X_L, X_{L-1}, \dots, X_1) : M \in b_{lyi} \otimes a_{kxh}\}$ where $X_j = W_{y \sim j}^\top M W_{j-1 \sim x}^\top$ subspace of $\mathbb{R}^{d_\theta}$

$\psi_{lkih} = \psi_{lklhih} = \{(X_L, X_{L-1}, \dots, X_1) : M \in b_{lli} \otimes a_{khh}\}$ where $X_j = W_{l \sim j}^\top M W_{j-1 \sim h}^\top$ subspace of $\mathbb{R}^{d_\theta}$

$\Psi^{\text{free}} = \{\psi_{Lki0} \neq \{\mathbf{0}\} : k, i \in [0, L] \text{ and } k + 1 \geq i\}$ freedom normal prebasis; spans row $d\mu(\theta)$

$\Psi^{\text{stratum}} = \Psi^{\text{free}} \cup \{\psi_{lkih} \neq \{\mathbf{0}\} : L \geq l \geq k + 1 \geq i > h \geq 0\}$ stratum normal prebasis; spans $N_\theta S$

$\omega_{ki} = \dim a_{kji} = \dim b_{kji} = \text{rk } W_{k \sim i} - \text{rk } W_{k \sim i-1} - \text{rk } W_{k+1 \sim i} + \text{rk } W_{k+1 \sim i-1}$ interval $[i, k]$ multiplicity

---

$\underline{r} \leq \underline{s}$ $r_{k \sim i} \leq s_{k \sim i}$ for all $L \geq k \geq i \geq 0$

$\underline{r} < \underline{s}$ $\underline{r} \leq \underline{s}$ and $\underline{r} \neq \underline{s}$ (at least one of the inequalities holds strictly)

$\overline{S}$ closure of set $S \subseteq \mathbb{R}^{d_\theta}$ with respect to weight space $\mathbb{R}^{d_\theta}$

$Z^\perp$ orthogonal complement of subspace Z (subspace of all vectors orthogonal to all vectors in Z)

$\text{proj}_Z Y = Z \cap (Z^\perp + Y)$ orthogonal projection of subspace Y onto subspace Z

$MZ = \{Mv : v \in Z\}$ product of a matrix M and a subspace Z

$Z = X \oplus Y$ direct sum decomposition: $Z = X + Y$ (vector sum of subspaces) and $X \cap Y = \{\mathbf{0}\}$

$Z = \bigoplus_{i=1}^{m} X_i$ direct sum decomposition: $Z = \sum_{i=1}^{m} X_i$ and for every $i \in [1, m]$, $X_i \cap \sum_{j \neq i} X_j = \{\mathbf{0}\}$

$Z \downarrow Y = \{X \subseteq Z : Z = X \oplus Y\}$ subspaces of Z of dimension $\dim Z - \dim Y$ linearly independent of Y

$U \otimes V = \{M \in \mathbb{R}^{p \times q} : \text{col } M \subseteq U \text{ and row } M \subseteq V\}$ tensor product of subspaces $U \subseteq \mathbb{R}^p$ and $V \subseteq \mathbb{R}^q$

$\text{col } M$; $\text{col } d\mu(\theta)$ columnspace of matrix M or of linear map $d\mu(\theta)$

$\dim S$ dimension of subspace or manifold S

$\text{null } M$; $\text{null } d\mu(\theta)$ nullspace of matrix M or of linear map $d\mu(\theta)$

$\text{null } M^\top$; $\text{null } d\mu^\top(\theta)$ left nullspace of matrix M or of linear map $d\mu(\theta)$

$\text{rk } M = \dim \text{col } M = \dim \text{row } M$; $\text{rk } d\mu(\theta) = \dim \text{col } d\mu(\theta) = \dim \text{row } d\mu(\theta)$ rank of M or of $d\mu(\theta)$

$\text{row } M$; $\text{row } d\mu(\theta)$ rowspace of matrix M or of linear map $d\mu(\theta)$

$\text{span } \Theta = \sum_{\sigma \in \Theta} \sigma$ vector sum of the subspaces in a prebasis Θ

Table 2: More notation used in this paper. (Continued from Table 1.) See also Table 6.

3   A Foretaste of our Results

The fiber $\mu^{-1}(W)$ is a real algebraic variety (again, the set of all real solutions to a system of polynomial equations—in our setting, multilinear equations). While some fibers of the matrix multiplication map are manifolds, as Figure 1 illustrates, in general they are not manifolds, as Figure 2 illustrates. In 1957, Hassler Whitney [22–24] proved that every algebraic variety can be partitioned into a set of smooth manifolds (not necessarily closed manifolds) of varying dimensions. These manifolds are called *strata*, and the set of manifolds is called a *stratification*. By "partitioned," we mean that the strata are pairwise disjoint and the fiber is the union of the strata. By "smooth," we mean that there exists a stratification[1] whose strata are differentiable manifolds of class C^∞. Figure 2 depicts a stratification of a fiber with three strata.

In this paper, we show that the fiber $\mu^{-1}(W)$ of a matrix W under the matrix multiplication map μ has a natural stratification by rank list. The *rank stratification* is

$$\mathcal{S} = \{S_{\underline{r}} : \underline{r} \text{ is the rank list of some weight vector in } \mu^{-1}(W)\}.$$

It is clear that the members of $\mathcal{S}$ are disjoint and that $\mu^{-1}(W) = \bigcup_{S \in \mathcal{S}} S$—that is, $\mathcal{S}$ is a partition of $\mu^{-1}(W)$. One of our main results (Theorem 16) is that each $S_{\underline{r}}$ is a C^∞-differentiable manifold (without boundary, but not necessarily closed nor connected nor bounded).[2] As each $S_{\underline{r}}$ is a manifold, $\mathcal{S}$ is a stratification and each $S_{\underline{r}}$ in $\mathcal{S}$ is a stratum. $\mathcal{S}$ contains finitely many strata—one stratum for each rank list that occurs among the fiber's points. We will study the properties of the rank stratification and use it to understand the geometry and topology of the fiber.

Recall that $\bar{S}$ denotes the closure of S. A stratification satisfies the *frontier condition* if for every pair of distinct strata S, T in the stratification, either $S \cap \bar{T} = \emptyset$ or $S \subseteq \bar{T}$. (Both statements are true if $S = \emptyset$; otherwise, they are mutually exclusive.) That is, the inclusion of points of S in $\bar{T}$ is all or nothing, which makes it easier to understand how strata are connected to each other. For example, if $S \subseteq \bar{T}$ and $S \neq T$, then $\dim S < \dim T$ and from any point on S there is an infinitesimal perturbation that takes us onto T. We will show that rank stratifications satisfy the frontier condition (Theorem 27).

The relationship $S \subseteq \bar{T}$ is transitive: if $S \subseteq \bar{T}$ and $T \subseteq \bar{U}$, then $S \subseteq \bar{U}$. Hence it induces a partial ordering of the strata in the rank stratification. This partial ordering is easily inferred from the rank lists. Consider two strata $S_{\underline{r}} \neq \emptyset$ and $S_{\underline{s}} \neq \emptyset$ from the same fiber, with rank lists $\underline{r}$ and $\underline{s}$. Recall that $\underline{r} \leq \underline{s}$ means that $r_{k \sim i} \leq s_{k \sim i}$ for all $L \geq k \geq i \geq 0$. We will show that $S_{\underline{r}} \subseteq \bar{S}_{\underline{s}}$ if and only if $\underline{r} \leq \underline{s}$ (Theorem 27). It is easy to show that $S_{\underline{r}} \subseteq \bar{S}_{\underline{s}}$ implies $\underline{r} \leq \underline{s}$, but proving the reverse implication required us to solve a tricky combinatorial puzzle.

Return to Figure 2, which graphs the variety of solutions to $W_2 W_1 = [\theta_2][\theta_1 \quad \theta_1'] = [0 \quad 0] = W$ and depicts a stratification of that variety. There are two ways to achieve $W_2 W_1 = [0 \quad 0]$: 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. Recall that the subscripts of S_{**} are $\text{rk } W_2$ and $\text{rk } W_1$. If we set both W_2 and W_1 to zero, our

---

[1]In general, the term "stratification" does not require that the strata be smooth; only that they be manifolds. We can speak of a stratification of a topological space without considering any embedding of the space at all.

[2]The strata are, in the language of topology, *manifolds without boundary*, as for every stratum S and every point $\zeta \in S$, there is an open neighborhood $N \subset S$ that contains ζ 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 $\mathbb{R}^{d_\theta} \setminus S$. In our setting, 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 its closure $\bar{S}$. For example, in Figure 2, S_{01} is a plane with the origin removed, the closure of S_{01} is the whole plane, and the boundary of S_{01} also is the whole plane. So our manifolds without (intrinsic) boundary have (extrinsic) boundaries. Note that S_{01} is an example of a stratum that is neither closed nor bounded, and S_{10} is a stratum that is not closed, bounded, nor connected.
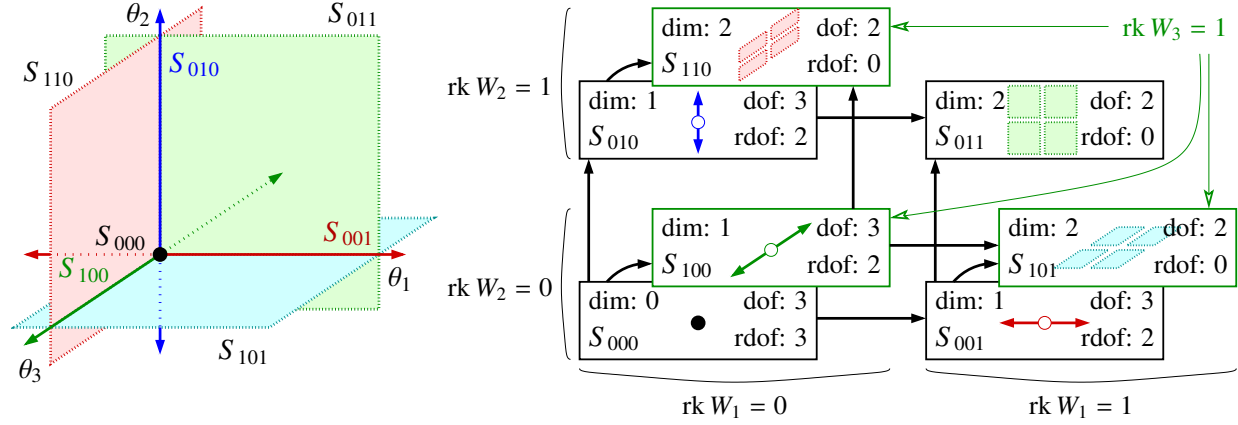
Figure 3: At left is the variety of solutions to $W_3 W_2 W_1 = [\theta_3][\theta_2][\theta_1] = [0] = W$, partitioned into seven strata: S_{000} is the origin; S_{001}, S_{010}, and S_{100} are the three coordinate axes with the origin removed; and S_{011}, S_{101}, and S_{110} are the three coordinate planes with the coordinate axes removed. At right is the stratum dag, organized as a three-dimensional table indexed by the ranks of W_3, W_2, and W_1.

weight vector $\theta = (W_2, W_1)$ is the origin, a 0-dimensional stratum labeled S_{00}. If we set only W_2 to zero, our weight vector lies on the stratum S_{01}, the pink plane with the origin removed. If we set only W_1 to zero, our weight vector lies on the stratum S_{10}, the blue line with the origin removed. Consistent with our claims above, $S_{00} \subset \bar{S}_{01}$ and $S_{00} \subset \bar{S}_{10}$, whereas $S_{10} \cap \bar{S}_{01} = \emptyset$ and $S_{01} \cap \bar{S}_{10} = \emptyset$.

We have omitted the other ranks in the rank list from the subscripts of S in Figure 2 because only rk W_2 and rk W_1 vary. The omitted ranks, the d_j's and rk W, are fixed for a specific fiber. In general, a rank list has $(L^2 + 3L)/2 + 1$ numbers, but if we leave out the d_j's and rk W, we have $(L^2 + L)/2 - 1$ numbers. Hence it is natural to organize the strata in a table or array with $(L^2 + L)/2 - 1$ indices or dimensions.

At right in Figure 2, we arrange the strata in a directed acyclic graph (dag) we call the *stratum dag*, which represents the stratification as a partially ordered set of strata. If the dag contains a directed path from S_r to S_s, then $S_r \subset \bar{S}_s$, dim $S_r <$ dim S_s, and $r < s$. (See Section 3.1 for an explanation of the edges of the dag.) For each stratum, the dag 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"), and thus represent moves off the stratum onto a higher-dimensional stratum. These numbers always satisfy dof = dim + rdof.

Figure 3 depicts another example of a nonmanifold fiber, whose rank stratification has seven strata. The fiber $\mu^{-1}([0])$ is the variety of solutions to $[\theta_3][\theta_2][\theta_1] = [0]$ (an instantiation of $W_3 W_2 W_1 = W$). We arrange the stratum dag in a three-dimensional table, 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 $W_3 W_2$ and rk $W_2 W_1$ can vary as well; but in this example every matrix is 1×1, so those two ranks are uniquely determined by the first three. Note that although we draw no directed edge from S_{000} directly to S_{011}, the presence of a directed path from S_{000} to S_{011} in the dag implies that $S_{000} \subset \bar{S}_{011}$. In this example, every stratum's closure includes S_{000}.

Figure 4 depicts the stratum dag representing the rank stratification of the fiber $\mu^{-1}(W)$ for any 5×4 matrix W with rank 1 and a network with $L = 2$, $d_2 = 5$, $d_1 = 6$, and $d_0 = 4$. The fiber has dimension as high as 35 at some points, and it is embedded in a 54-dimensional weight space. Unfortunately, at this size we cannot visualize the geometry of the fiber. But this example begins to give some insight into the structure of more complicated fibers. A notable aspect of this example is that it has many strata branching out from each other, like in Figure 3, but they are curved, like in Figure 1.

	rk $W_1 = 1$	rk $W_1 = 2$	rk $W_1 = 3$	rk $W_1 = 4$
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

Figure 4: Stratum dag representing the stratification of $\mu^{-1}(W)$ for $W = W_2 W_1$, $W_2 \in \mathbb{R}^{5\times 6}$, $W_1 \in \mathbb{R}^{6\times 4}$, 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. For every pair of strata S_{ki} and $S_{k'i'}$ with $k \leq k'$ and $i \leq i'$, $S_{ki} \subset \bar{S}_{k'i'}$.

One of our main results is that the dimension of a stratum in the rank stratification ("dim" in the stratum dags) is

$$D^{\text{stratum}} = d_\theta - \text{rk } W \cdot (d_L + d_0 - \text{rk } W) - \sum_{L \geq k+1 \geq i > 0} (\text{rk } W_{k+1\sim i} - \text{rk } W_{k+1\sim i-1})(\text{rk } W_{k\sim i-1} - \text{rk } W_{k+1\sim i-1}),$$

hence solely a function of its rank list. The derivation is not simple: in Section 8.7 we prove that D^{stratum} is a lower bound on the dimension of a stratum, and in Section 9.3 we prove that it is an upper bound.

We also prove that in the rank stratification, the topology of a stratum is determined solely by its rank list, and its geometry is determined solely by its rank list and a linear transformation in weight space. Specifically, if we consider two strata that come from two different fibers but have the same rank list, then there is a linear homeomorphism mapping one stratum to the other (Corollary 14). Moreover, there is a sense in which every point on a stratum looks the same: if we specify two points on the same stratum, there is a bijective linear transformation mapping the stratum to itself and mapping one specified point to the other (Corollary 15).

We devote a lot of writing to deriving the tangent and normal spaces at each point on a stratum. (Deriving these spaces is part of our derivation of the dimension of a stratum.) These spaces help us understand the geometry of the fiber. Table 3 tabulates the most important of these results, including explicit formulae for the space tangent to a stratum at a particular point, $T_\theta S$ (derived in Section 8.6), and the space normal to a stratum, $N_\theta S$ (derived in Section 9.3). While it is nice to be able to write explicit formulae in terms of nullspaces, rowspaces, and columnspaces, for practical purposes it is more useful to compute a basis for each of these spaces, and we devote even more writing to that task (Sections 7 through 9).

We are particularly interested in points on the fiber where multiple strata meet. Given a point (weight vector) $\theta \in \mu^{-1}(W)$, we describe a *fiber basis* composed of linearly independent vectors having the property that for every stratum S whose closure contains θ, some subset of the fiber basis spans the tangent space $T_\theta \bar{S}$. (This assumes that $\bar{S}$ has a tangent space at θ—if not, a more subtle characterization is needed.) Each vector in the fiber basis is tangent to some smooth path leaving θ on the fiber. For example, in Figure 3, at a point $\theta \in S_{001}$, the three unit coordinate vectors can serve as the fiber basis: the tangent plane $T_\theta \bar{S}_{011}$ is spanned by the θ_1- and θ_2-axes, and the tangent plane $T_\theta \bar{S}_{101}$ is spanned by the θ_1- and θ_3-axes. At a point $\theta \in S_{101}$, the fiber basis has just two vectors—the θ_1- and θ_3-axes will do.

$$
\begin{aligned}
\dim S \;=\;& d_\theta - \operatorname{rk} W \cdot (d_L + d_0 - \operatorname{rk} W) - \\
& \sum_{L \ge k+1 \ge i > 0} (\operatorname{rk} W_{k+1 \sim i} - \operatorname{rk} W_{k+1 \sim i-1})(\operatorname{rk} W_{k \sim i-1} - \operatorname{rk} W_{k+1 \sim i-1}).
\end{aligned}
$$

$$
\begin{aligned}
T_\theta S \;=\;& \Big\{ (\Delta W_L + W_L H_{L-1},\, \Delta W_{L-1} + W_{L-1}H_{L-2} - H_{L-1}W_{L-1}, \ldots, \\
& \Delta W_j + W_j H_{j-1} - H_j W_j, \ldots, \Delta W_2 + W_2 H_1 - H_2 W_2,\, \Delta W_1 - H_1 W_1) : \\
& H_j \in \mathbb{R}^{d_j \times d_j}, \\
& \Delta W_j \in \sum_{h=1}^{j-1} \operatorname{col} W_{j \sim h} \otimes \operatorname{null} W_{j-1 \sim h-1}^\top + (\operatorname{null} W_{L \sim j} \cap \operatorname{col} W_{j \sim 0}) \otimes \mathbb{R}^{d_{j-1}} + \\
& \sum_{l=j}^{L-1} \operatorname{null} W_{l+1 \sim j} \otimes \operatorname{row} W_{l \sim j-1} + \mathbb{R}^{d_j} \otimes (\operatorname{row} W_{L \sim j-1} \cap \operatorname{null} W_{j-1 \sim 0}^\top) \Big\}.
\end{aligned}
$$

$$
\dim T_\theta S \;=\; \dim S.
$$

$$
\begin{aligned}
N_\theta S \;=\;& \Big\{ (X_L^{L0}, X_{L-1}^{L0}, \ldots, X_1^{L0}) : M^{L0} \in \mathbb{R}^{d_L \times d_0} \Big\} + \\
& \sum_{L \ge y > x \ge 0} \Big\{ (X_L^{yx}, X_{L-1}^{yx}, \ldots, X_1^{yx}) : M^{yx} \in \operatorname{null} W_{y \sim x}^\top \otimes \operatorname{null} W_{y \sim x} \Big\} \qquad \text{where}
\end{aligned}
$$

$$
X_j^{yx} \;=\; \begin{cases} W_{y \sim j}^\top M^{yx} W_{j-1 \sim x}^\top, & j \in [x+1, y], \\ 0, & j \notin [x+1, y]. \end{cases}
$$

$$
\dim N_\theta S \;=\; d_\theta - \dim S.
$$

$$
\operatorname{rk} d\mu(\theta) \;=\; \sum_{j=1}^{L} \operatorname{rk} W_{L \sim j} \cdot \operatorname{rk} W_{j-1 \sim 0} - \sum_{j=1}^{L-1} \operatorname{rk} W_{L \sim j} \cdot \operatorname{rk} W_{j \sim 0} \qquad \text{(Trager et al. [21], Lemma 3)}
$$

$$
\begin{aligned}
\operatorname{null} d\mu(\theta) \;=\;& \Big\{ (\Delta W_L + W_L H_{L-1},\, \Delta W_{L-1} + W_{L-1}H_{L-2} - H_{L-1}W_{L-1}, \ldots, \\
& \Delta W_j + W_j H_{j-1} - H_j W_j, \ldots, \Delta W_2 + W_2 H_1 - H_2 W_2,\, \Delta W_1 - H_1 W_1) : \\
& H_j \in \mathbb{R}^{d_j \times d_j},\, \Delta W_j \in \operatorname{null} W_{L \sim j} \otimes \mathbb{R}^{d_{j-1}} + \mathbb{R}^{d_j} \otimes \operatorname{null} W_{j-1 \sim 0}^\top \Big\}.
\end{aligned}
$$

$$
\dim \operatorname{null} d\mu(\theta) \;=\; d_\theta - \operatorname{rk} d\mu(\theta).
$$

$$
\operatorname{row} d\mu(\theta) \;=\; \{ (X_L, X_{L-1}, \ldots, X_1) : M \in \mathbb{R}^{d_L \times d_0} \} \qquad \text{where}
$$

$$
X_j \;=\; W_{L \sim j}^\top M W_{j-1 \sim 0}^\top.
$$

$$
\dim \operatorname{row} d\mu(\theta) \;=\; \operatorname{rk} d\mu(\theta).
$$

Notes: $N_\theta S = (T_\theta S)^\perp$; $\operatorname{row} d\mu(\theta) = (\operatorname{null} d\mu(\theta))^\perp$;
$T_\theta S \subseteq \operatorname{null} d\mu(\theta)$; $N_\theta S \supseteq \operatorname{row} d\mu(\theta)$.

Table 3: Some of this paper's main results about subspaces of the weight space $\mathbb{R}^{d_\theta}$ and their dimensions. S is the stratum of the fiber $\mu^{-1}(W)$ containing the point $\theta = (W_L, W_{L-1}, \ldots, W_1)$, $T_\theta S$ is the tangent space of S at θ ($T_\theta S$ passes through the origin, not through θ), $N_\theta S$ is the normal space of S at θ (the orthogonal complement of $T_\theta S$), and $d\mu(\theta)$ is the (linear) differential map of the matrix multiplication map μ at θ, defined in Section 8.

The example in Figure 3 is a bit misleading. In general, the vectors in the fiber basis cannot always be orthogonal to each other, because strata do not always meet each other at right angles. Moreover, every point on $\mu^{-1}(W)$ may need a different fiber basis, because usually fibers are curved. Even the size of the fiber basis is different at different points on the fiber: the "degrees of freedom" (dof) in our stratum dags is the number of vectors in the fiber basis at a specific point θ. But at any one point $\theta \in \mu^{-1}(W)$, a single fiber basis suffices to describe all the tangencies, so we gain an understanding of *how* the strata meet at θ.

3.1 Some Details about Stratum Dags

One goal of this paper is to provide an algorithm that generates stratum dags for any $L \geq 2$, given the d_j's and rk W as input. That algorithm appears in Section 10. Here we discuss some finer points of stratum dags and their structure.

Figures 2, 3, and 4 illustrate that some fibers have rank stratifications with multiple maximal elements—that is, there is no single stratum whose closure includes the whole fiber (equivalently, no dag vertex that is reachable from all the other vertices). By contrast, every fiber's rank stratification has one unique minimal element (equivalently, a dag vertex from which all the other vertices are reachable): the stratum for which every rank (except the d_j's) is equal to rk W. For instance, in Figure 4, $S_{11} \subseteq \bar{S}_{ki}$ for every stratum S_{ki}.

For two-matrix fibers ($L = 2$), the general table shape is a pentagon like the one illustrated in Figure 4. We have drawn the stratum dag so that the horizontal axis specifies rk W_1 and the vertical axis specifies rk W_2 for each stratum. 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: by Sylvester's inequality, rk $W_2 + \text{rk } W_1 \leq d_1 + \text{rk } W$. In Figure 4, that means rk $W_2 + \text{rk } W_1 \leq 7$. This inequality generates the fifth edge of the pentagon.

When L is large, the shape of the table is more complicated because then we have not one inequality, but many occurrences of the Frobenius rank inequality dictating how the ranks in the rank list constrain each other. For example, rk $W_3 W_2$ cannot vary entirely independently of rk W_3 and rk W_2. (See Section 10.1 for details.) Figures 2, 3, and 4 might give the false impression that stratum dag edges are always axis-aligned, representing a change of a single rank. But in networks with more layers, the stratum dags typically have edges that represent the increase of multiple ranks simultaneously. (See Section 10.2, and also Figure 9.)

Which strata are connected by edges of a stratum dag? Unfortunately, we haven't yet covered the background to answer that question clearly, but: the dag has a directed edge $(S_{\underline{r}}, S_{\underline{s}})$ if there exists a *rank-1 abstract connecting or swapping move*, described in Section 7.5, that transforms the rank list $\underline{r}$ to the rank list $\underline{s}$. Connecting and swapping moves are of great interest to us, because they codify the relationship between the rank lists of strata $S_{\underline{t}}$ and $\bar{S}_{\underline{u}}$ satisfying $\underline{t} < \underline{u}$: there is a sequence of abstract connecting and swapping moves that transform $\underline{t}$ to $\underline{u}$ (represented by a directed path in the stratum dag).

The stratum dag is not necessarily the simplest dag that represents the partial ordering of the strata (sometimes called a *Hasse diagram*[3]), because sometimes a stratum dag contains a directed edge $(S_{\underline{r}}, S_{\underline{s}})$ despite also containing the edges $(S_{\underline{r}}, S_{\underline{t}})$ and $(S_{\underline{t}}, S_{\underline{s}})$. Section 7.5 and Figure 11 give an example of that. Roughly speaking, an edge $(S_{\underline{r}}, S_{\underline{s}})$ indicates that from a weight vector $\theta \in S_{\underline{r}}$, the stratum $S_{\underline{s}}$ permits one or more degrees of freedom of motion not "covered" by any of the strata $S_{\underline{t}}$ such that $\underline{r} \leq \underline{t} < \underline{s}$ (i.e., the predecessors of $S_{\underline{s}}$ in the stratum dag whose closures contain θ). Contrast Figure 11 with Figure 3, in which the stratum dag has no edge (S_{000}, S_{011}) because, although there are two degrees of freedom by which the weight vector in S_{000} can be perturbed onto S_{011}, they are spanned by the degree of freedom of motion on S_{001} and the degree of freedom of motion on S_{010}.

---

[3]To use some more precise terms: the stratum dag is not necessarily its own transitive reduction; the edges of the stratum dag do not necessarily express the covering relation of the partial order.

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 hidden 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 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 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 illuminates gradient descent algorithms for learning, whose success may depend on injecting information into these channels.

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 W's rank above 1. (In the language of Trager et al. [21], the cost function has a spurious critical point.) 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 rank of W.

The main theoretical motivation arises because the rank stratification of the fiber $\mu^{-1}(W)$ of a matrix W has one stratum (and one rank list) for each specific state of subspace flow through the network. Moreover, we will use the subspaces we discover to derive the tangent and normal spaces of the strata. Thus these different "states of subspace flow" help us to understand the topology and geometry of the fiber. These states 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 the 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 ($\omega_{ki} = 0$).

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
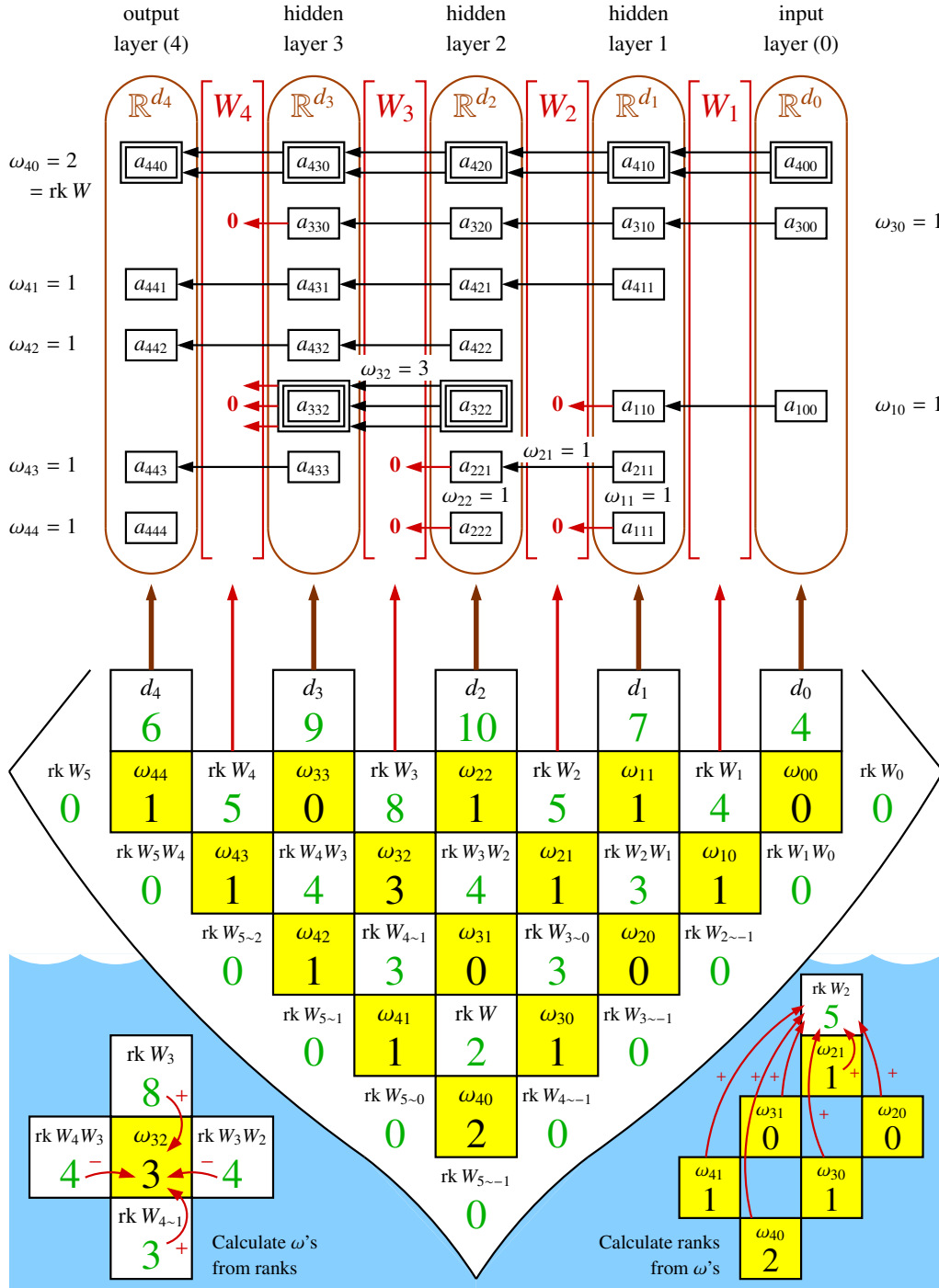
Figure 5: The tea clipper ship Basis Flow. The top half is a basis flow diagram that illustrates the flow of the prebasis subspaces a_{kji} through the network. Double boxes represent subspaces of dimension 2 and triple boxes represent subspaces of dimension 3. The bottom half shows the relationships 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_{ts} of the intervals that touch layer j (i.e., the dimensions of the prebasis subspaces a_{tjs}). Each matrix rank rk $W_{k\sim i}$ is the sum of the multiplicities of the intervals that touch both layers k and i.

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).

The simplest example is a one-matrix network ($L = 1$), for which the subspaces we speak of are the four fundamental subspaces row W_1, col W_1, null W_1, and null $W_1^{\top}$. The interval $[0, 1]$ represents the rowspace of W_1 (at the input layer) and the fact that applying W_1 to row W_1 yields the columnspace of W_1 (at the output layer). Both row W_1 and col W_1 have dimension $\omega_{10} = \text{rk } W_1$. The interval $[0, 0]$ represents the nullspace of W_1, which disappears into W_1, sending no information to the output layer. The dimension of null W_1 is ω_{00}. The interval $[1, 1]$ represents the left nullspace of W_1, the unused dimensions of the output layer. The dimension of null $W_1^{\top}$ is ω_{11}.

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 derive tangent and normal spaces on the fiber $\mu^{-1}(W)$.

A multiset of intervals is fully specified by the parameters $\omega_{ki} \geq 0$ for all k and i satisfying $L \geq k \geq i \geq 0$. 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,

$$d_j = \sum_{t=j}^{L} \sum_{s=0}^{j} \omega_{ts} \quad \forall j \in [0, L]. \tag{4.1}$$

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. Therefore, only finitely many valid multisets are possible for a network with fixed layer sizes d_j. The constraint (4.1) reflects that we will build a basis for each layer of units: 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 $\mathbb{R}^{d_j}$.

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,

$$\text{rk } W_{k \sim i} = \sum_{t=k}^{L} \sum_{s=0}^{i} \omega_{ts}. \tag{4.2}$$

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

In particular,

$$\text{rk } W = \omega_{L0}. \tag{4.3}$$

That is, the interval $[0, L]$ always has multiplicity rk W. This interval represents the fact that the subspace row W at the input layer is mapped by W to col W at the output layer, and both subspaces have dimension rk W. For a specific matrix W, the rank rk W and the multiplicity ω_{L0} are fixed, as is $d_j = \text{rk } W_{j \sim j}$ for $j \in [0, L]$. The other ranks and multiplicities may vary across different factorizations of W.

Each valid multiset is associated with a particular rank list, a particular basis flow diagram like Figure 5, and a particular stratum of the fiber $\mu^{-1}(W)$ for any $W \in \mathbb{R}^{d_L \times d_0}$ with rank ω_{L0}. A rank list $\underline{r}$ is *valid* if there is some weight vector $\theta \in \mathbb{R}^{d_\theta}$ that has rank list $\underline{r}$. In Section 4.5 we show that there is a bijection from valid rank lists to valid multisets of intervals: if we are given a rank list, we can easily determine the interval multiplicities, and if we are given a list of interval multiplicities, we can easily determine the ranks. (In Appendix A we elaborate on why every valid multiset of intervals maps to a valid rank list.)

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.

$$\mathbb{R}^{d_j} = \text{row } W_{j\sim j} \supseteq \text{row } W_{j+1} \supseteq \text{row } W_{j+2}W_{j+1} \supseteq \ldots \supseteq \text{row } W_{L\sim j} \supseteq \text{row } W_{L+1\sim j} = \{\mathbf{0}\},$$

$$\{\mathbf{0}\} = \text{null } W_{j\sim j} \subseteq \text{null } W_{j+1} \subseteq \text{null } W_{j+2}W_{j+1} \subseteq \ldots \subseteq \text{null } W_{L\sim j} \subseteq \text{null } W_{L+1\sim j} = \mathbb{R}^{d_j},$$

$$\mathbb{R}^{d_j} = \text{col } W_{j\sim j} \supseteq \text{col } W_j \supseteq \text{col } W_j W_{j-1} \supseteq \ldots \supseteq \text{col } W_{j\sim 0} \supseteq \text{col } W_{j\sim -1} = \{\mathbf{0}\}, \quad \text{and}$$

$$\{\mathbf{0}\} = \text{null } W_{j\sim j}^\top \subseteq \text{null } W_j^\top \subseteq \text{null } (W_j W_{j-1})^\top \subseteq \ldots \subseteq \text{null } W_{j\sim 0}^\top \subseteq \text{null } W_{j\sim -1}^\top = \mathbb{R}^{d_j}.$$

By the Fundamental Theorem of Linear Algebra, 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.

$$W_{j\sim j} = I_{d_j \times d_j} \quad \text{(the } d_j \times d_j \text{ identity matrix)}.$$

Hence, row $W_{j\sim j} = \mathbb{R}^{d_j} = \text{col } W_{j\sim j}$ and null $W_{j\sim j} = \{\mathbf{0}\} = \text{null } W_{j\sim j}^\top$.

$$W_{j\sim -1} = 0_{d_j \times 1} \quad \text{and} \quad W_{L+1\sim j} = 0_{1 \times d_j} \quad \text{(zero matrices)}.$$

Hence, row $W_{L+1\sim j} = \{\mathbf{0}\} = \text{col } W_{j\sim -1}$ and null $W_{L+1\sim j} = \mathbb{R}^{d_j} = \text{null } W_{j\sim -1}^\top$.

(Note that the last two lines are consistent with imagining that the network $W_L W_{L-1} \cdots W_1$ 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

$$A_{kji} = \text{null } W_{k+1\sim j} \cap \text{col } W_{j\sim i}, \quad k \in [j-1, L], i \in [-1, j], \quad \text{and} \tag{4.4}$$

$$B_{kji} = \text{row } W_{k\sim j} \cap \text{null } W_{j\sim i-1}^\top, \quad k \in [j, L+1], i \in [0, j+1]. \tag{4.5}$$

For example, $A_{320} = \text{null } W_4 W_3 \cap \text{col } W_2 W_1$ and $B_{320} = \text{row } W_3 \cap \text{null } W_{2\sim -1}^\top = \text{row } W_3$. 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

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

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

		$i = -1$		$i = 0$		$i = 1$		$i = 2$
$\mathbb{R}^{d_2}$	$k = 4$	$A_{4,2,-1} = \{\mathbf{0}\}$	$\subseteq$	A_{420}	$\subseteq$	A_{421}	$\subseteq$	$A_{422} = \mathbb{R}^{d_2}$
$\cup\mathsf{I}$				$\cup\mathsf{I}$		$\cup\mathsf{I}$		$\cup\mathsf{I}$
null W_4W_3	$k = 3$	$A_{3,2,-1} = \{\mathbf{0}\}$	$\subseteq$	A_{320}	$\subseteq$	A_{321}	$\subseteq$	A_{322}
$\cup\mathsf{I}$				$\cup\mathsf{I}$		$\cup\mathsf{I}$		$\cup\mathsf{I}$
null W_3	$k = 2$	$A_{2,2,-1} = \{\mathbf{0}\}$	$\subseteq$	A_{220}	$\subseteq$	A_{221}	$\subseteq$	A_{222}
$\cup\mathsf{I}$				$\cup\mathsf{I}$		$\cup\mathsf{I}$		$\cup\mathsf{I}$
$\{\mathbf{0}\}$	$k = 1$	$A_{1,2,-1} = \{\mathbf{0}\}$		$A_{120} = \{\mathbf{0}\}$		$A_{121} = \{\mathbf{0}\}$		$A_{122} = \{\mathbf{0}\}$
null $W_{k+1\sim2}$		$i = -1$		$i = 0$		$i = 1$		$i = 2$
$A_{k2i} \nearrow$	col $W_{2\sim i}$	$\{\mathbf{0}\}$	$\subseteq$	col W_2W_1	$\subseteq$	col W_2	$\subseteq$	$\mathbb{R}^{d_2}$

		$i = 0$		$i = 1$		$i = 2$		$i = 3$
$\{\mathbf{0}\}$	$k = 5$	$B_{520} = \{\mathbf{0}\}$		$B_{521} = \{\mathbf{0}\}$		$B_{522} = \{\mathbf{0}\}$		$B_{523} = \{\mathbf{0}\}$
$\cap\mathsf{I}$		$\cap\mathsf{I}$		$\cap\mathsf{I}$		$\cap\mathsf{I}$		
row W_4W_3	$k = 4$	B_{420}	$\supseteq$	B_{421}	$\supseteq$	B_{422}	$\supseteq$	$B_{423} = \{\mathbf{0}\}$
$\cap\mathsf{I}$		$\cap\mathsf{I}$		$\cap\mathsf{I}$		$\cap\mathsf{I}$		
row W_3	$k = 3$	B_{320}	$\supseteq$	B_{321}	$\supseteq$	B_{322}	$\supseteq$	$B_{323} = \{\mathbf{0}\}$
$\cap\mathsf{I}$		$\cap\mathsf{I}$		$\cap\mathsf{I}$		$\cap\mathsf{I}$		
$\mathbb{R}^{d_2}$	$k = 2$	$B_{220} = \mathbb{R}^{d_2}$	$\supseteq$	B_{221}	$\supseteq$	B_{222}	$\supseteq$	$B_{223} = \{\mathbf{0}\}$
row $W_{k\sim2}$		$i = 0$		$i = 1$		$i = 2$		$i = 3$
$B_{k2i} \nearrow$	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 4: The hierarchical nesting of the flow subspaces at unit layer $j = 2$ of a network with $L = 4$ matrices. Top: $A_{k2i} = $ null $W_{k+1\sim2} \cap$ col $W_{2\sim i}$ for each k, i. Bottom: $B_{k2i} = $ row $W_{k\sim2} \cap$ null $W_{2\sim i-1}$ for each k, i.

Table 4 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. Recall from Section 2 that 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, $B_{kji} = W_{y\sim j}^{\top} B_{kyi}$ 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} = $ null $W_{k+1\sim i} \cap \mathbb{R}^{d_i} = $ 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 $ null $W_{k+1\sim z}$. Obviously, $W_{z\sim i} A_{kii} \subseteq $ col $W_{z\sim i}$. Hence $W_{z\sim i} A_{kii} \subseteq $ null $W_{k+1\sim z} \cap $ col $W_{z\sim i} = A_{kzi}$.

We now show that the reverse inclusion also holds: $W_{z\sim i} A_{kii} \supseteq A_{kzi}$. Consider a vector $v \in A_{kzi}$. As $v \in $ 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 $ 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 $ 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 $z \in [i, k+1]$.
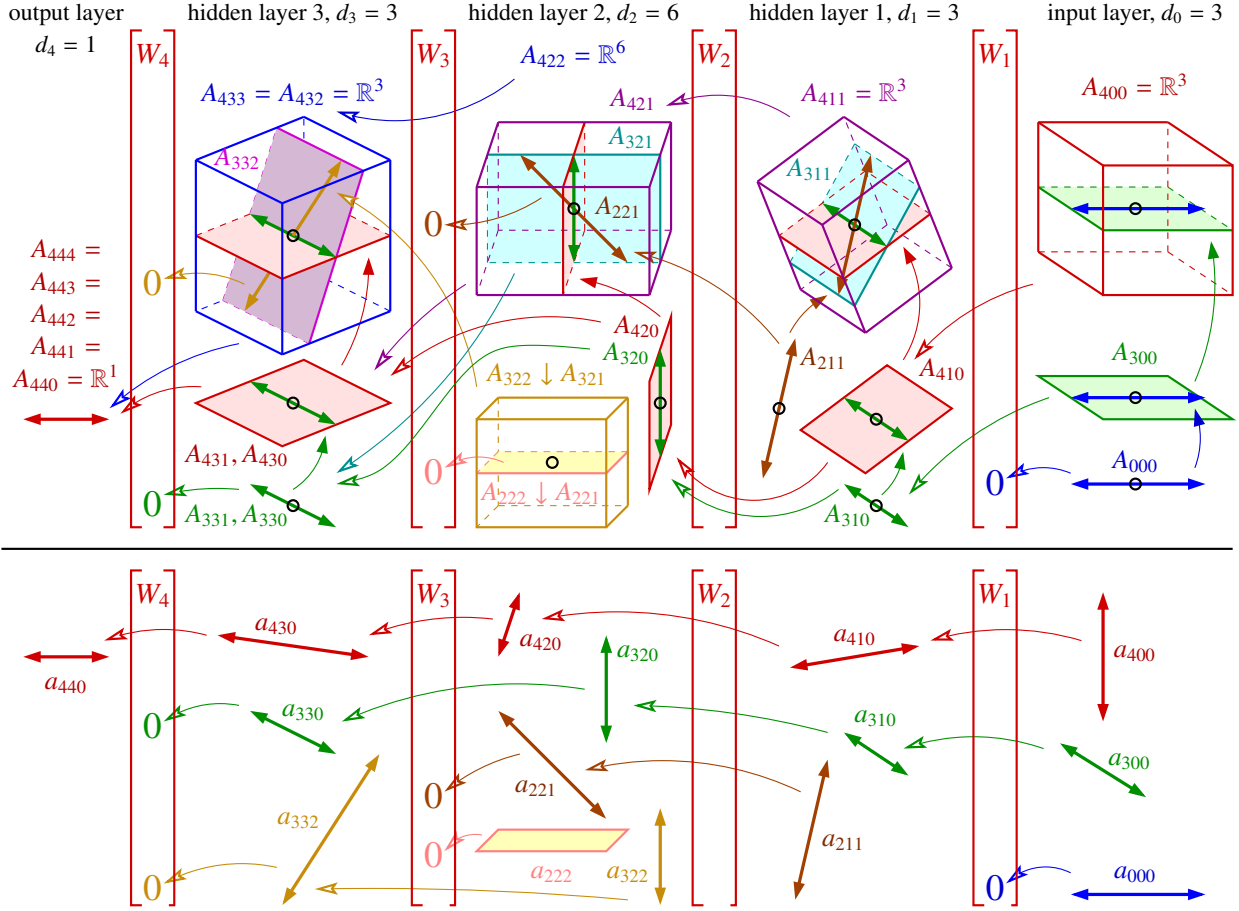
Figure 6: Top: an example of flow subspaces A_{kji}. Observe that A_{000} is annihilated in the nullspace of W_1, whereas A_{300} is not entirely annihilated until it reaches the nullspace of W_4. Observe that A_{410} and A_{311} meet at an oblique angle; flow subspaces do not always meet orthogonally. Note that the subspace A_{322} is five-dimensional, so we cannot draw it complete. Instead, we draw a three-dimensional subspace of A_{322} labeled $A_{322} \downarrow A_{321}$ such that A_{322} is the vector sum of the plane A_{321} and the space labeled $A_{322} \downarrow A_{321}$. Similarly, A_{222} is a three-dimensional subspace, but we draw a two-dimensional subspace of A_{222} labeled $A_{222} \downarrow A_{221}$ such that A_{222} is the vector sum of the line A_{221} and the plane labeled $A_{222} \downarrow A_{221}$. Bottom: an example of corresponding prebasis subspaces a_{kji}, forming flow prebases. The prebasis subspaces for layer j span $\mathbb{R}^{d_j}$.

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 $B_{kji} = W_{y\sim j}^{\top} B_{kyi}$. $\qquad\square$

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 layer j of 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 $X \oplus Y$.[4] Likewise, given a set of subspaces $X = \{X_1, X_2, \ldots, X_m\}$, the direct sum notation $X_1 \oplus X_2 \oplus \ldots \oplus X_m$ implies that the subspaces in 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 $X = \{X_1, X_2, \ldots, X_m\}$ is known as a *direct sum decomposition* of $\mathbb{R}^d$. That's too many syllables, so we will call 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 one and 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 $d = \dim X_1 + \dim X_2 + \ldots + \dim X_m$. A prebasis subspace X_i is a multidimensional analog of a basis vector. If desired, it is conceptually easy to convert a prebasis into a traditional 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." We will do that in Section 4.6, but we delay that step because details like the choice of basis for each prebasis subspace and the length of each basis vector are irrelevant here 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

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

$Z \downarrow Y$ contains the subspaces of Z that have dimension $\dim Z - \dim Y$ and are linearly independent of Y. In other words, $Z \downarrow Y$ is the set of all subspaces X yielding a direct sum $Z = X \oplus Y$ (i.e., $X \cap Y = \{\mathbf{0}\}$). 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, $Z \downarrow Y$ contains infinitely many 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 indices k, j, and i 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.

---

[4]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.

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. The bottom half of the figure would not change. 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: $(\text{row } W)^\perp = \text{null } W$ and $(\text{col } W)^\perp = \text{null } W^\top$. 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*

$$\dim(K \cap Z) - \dim(K \cap Y + J \cap Z)$$
$$= \dim(J^\perp \cap Y^\perp) - \dim(K^\perp \cap Y^\perp + J^\perp \cap Z^\perp)$$
$$= \dim(K \cap Z) - \dim(K \cap Y) - \dim(J \cap Z) + \dim(J \cap Y)$$
$$= \dim(J^\perp \cap Y^\perp) - \dim(K^\perp \cap Y^\perp) - \dim(J^\perp \cap Z^\perp) + \dim(K^\perp \cap Z^\perp).$$

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 De Morgan's laws $(E + F)^\perp = E^\perp \cap F^\perp$ and $(E \cap F)^\perp = E^\perp + F^\perp$.

$$\dim(J^\perp \cap Y^\perp) - \dim(K^\perp \cap Y^\perp) - \dim(J^\perp \cap Z^\perp) + \dim(K^\perp \cap Z^\perp)$$
$$= \dim J^\perp + \dim Y^\perp - \dim(J^\perp + Y^\perp) - \dim K^\perp - \dim Y^\perp + \dim(K^\perp + Y^\perp)$$
$$\quad - \dim J^\perp - \dim Z^\perp + \dim(J^\perp + Z^\perp) + \dim(K^\perp \cap Z^\perp)$$
$$= -\dim(J \cap Y)^\perp - \dim K^\perp + \dim(K \cap Y)^\perp - \dim Z^\perp + \dim(J \cap Z)^\perp + \dim(K + Z)^\perp$$
$$= -d + \dim(J \cap Y) - d + \dim K + d - \dim(K \cap Y) - d + \dim Z + d - \dim(J \cap Z) + d - \dim(K + Z)$$
$$= \dim(K \cap Z) - \dim(K \cap Y) - \dim(J \cap Z) + \dim(J \cap Y).$$

$\square$

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 $K = \text{null } W_{k+1 \sim j}$, $J = \text{null } W_{k \sim j}$, $Z = \text{col } W_{j \sim i}$, and $Y = \text{col } W_{j \sim i-1}$. (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 $B_{k+1,j,i+1} = K^\perp \cap Z^\perp$.) $\square$

We define prebases that span the subspaces A_{kji} and B_{kji}. Let

$$\mathcal{A}_{kji} = \{a_{k'ji'} \neq \{\mathbf{0}\} : k' \in [j,k], i' \in [0,i]\} \quad \text{and}$$
$$\mathcal{B}_{kji} = \{b_{k'ji'} \neq \{\mathbf{0}\} : k' \in [k,L], i' \in [i,j]\}.$$

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}. That is,*

$$A_{kji} = \bigoplus_{k' \in [j,k]} \bigoplus_{i' \in [0,i]} a_{k'ji'} \quad \text{and} \quad B_{kji} = \bigoplus_{k' \in [k,L]} \bigoplus_{i' \in [i,j]} b_{k'ji'}.$$

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, $w \in A_{k,j,i-1} \cap A_{k-1,j,i} = \text{null } W_{k\sim j} \cap \text{col } W_{j\sim i-1} = A_{k-1,j,i-1}$. 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 col $W_{j\sim i}$ (because $A_{Lji} = \text{col } W_{j\sim i}$) and $\mathcal{A}_{kjj}$ is a prebasis for null $W_{k+1\sim j}$ (because $A_{kjj} = \text{null } W_{k+1\sim j}$). 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 row $W_{k\sim j}$, and $\mathcal{B}_{jji}$ is a prebasis for 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 null $W_{k+1\sim j}$ and some 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 it is satisfying to explicitly write out the most natural candidates, akin to the rowspace, the nullspace, the columnspace, and the left nullspace of a matrix. Given two subspaces $S, T \in \mathbb{R}^d$, let $\text{proj}_S T$ 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*

$$
\begin{aligned}
a_{kji} &= \text{proj}_{\text{col } W_{j \sim i}} \text{row } W_{k \sim j} \cap \text{proj}_{\text{null } W_{k+1 \sim j}} \text{null } W_{j \sim i-1}^\top \\
&= \text{col } W_{j \sim i} \cap (\text{row } W_{k \sim j} + \text{null } W_{j \sim i}^\top) \cap \text{null } W_{k+1 \sim j} \cap (\text{row } W_{k+1 \sim j} + \text{null } W_{j \sim i-1}^\top) \quad and \quad (4.6) \\
b_{kji} &= \text{proj}_{\text{row } W_{k \sim j}} \text{col } W_{j \sim i} \cap \text{proj}_{\text{null } W_{j \sim i-1}^\top} \text{null } W_{k+1 \sim j} \\
&= \text{row } W_{k \sim j} \cap (\text{null } W_{k \sim j} + \text{col } W_{j \sim i}) \cap \text{null } W_{j \sim i-1}^\top \cap (\text{null } W_{k+1 \sim j} + \text{col } W_{j \sim i-1}). \quad (4.7)
\end{aligned}
$$

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, $Z \cap (Z \cap Y)^\perp = Z \cap (Z^\perp + Y^\perp) = \text{proj}_Z Y^\perp$. Hence

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

The third line implies (4.6). Symmetrically,

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

The third line implies (4.7). □

See Appendix B for additional discussion of the standard prebases.

4.4 Constructing Prebases 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*, which satisfy $a_{kji} = W_j a_{k,j-1,i}$ when $k \geq j > i$. Flow prebasis subspaces still have some flexibility: for instance, for $j \in [i + 1, k]$ we can choose a_{kii} and b_{kki} arbitrarily (for example, we could choose (4.6) for a_{kii} and (4.7) for b_{kki}), then obtain all the other subspaces by setting $a_{kji} = W_j a_{k,j-1,i}$ and $b_{k,j-1,i} = W_j^\top b_{kji}$. These subspaces flow through the linear neural network from specific starting layers to specific stopping layers, as expressed by a basis flow diagram such as Figure 5 (top) or Figure 6 (bottom), thereby outlining how information propagates (or would propagate, if it was there). 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} = $ null $W_{k\sim x} \cap$ col $W_{x\sim i}$. (That is, $a_{kxi} \cap A_{k-1,x,i} = \{\mathbf{0}\}$.) But $a_{kxi} \subseteq A_{kxi} \subseteq$ col $W_{x\sim i}$. Hence, every nonzero vector in a_{kxi} is in col $W_{x\sim i}$ but not in null $W_{k\sim x} \cap$ col $W_{x\sim i}$; thus no nonzero vector in a_{kxi} is in null $W_{k\sim x}$; thus no nonzero vector in a_{kxi} is in 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 $u-s-t \in$ null $W_{k\sim x}$. Recall that $A_{k-1,x,i} = $ null $W_{k\sim x} \cap$ col $W_{x\sim i}$. So $t \in$ null $W_{k\sim x}$, hence $u-s \in$ null $W_{k\sim x}$. Moreover, u and s are both in col $W_{x\sim i}$, so $u-s \in$ null $W_{k\sim x} \cap$ 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 $W_{y\sim j}^{\top}b_{kyi} \in B_{kji} \downarrow (B_{k,j,i+1} + B_{k+1,j,i})$ 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}.   □

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, as A_{311} meets A_{410} in Figure 6).

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, given fixed layer sizes $d_0, d_1, \ldots, d_L$, 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 index j as we now name this dimension.

Let

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

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 subspaces $a_{k,i+1,i}, a_{k,i+2,i}, \ldots$, 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 fibers $\mu^{-1}(W)$ and $\mu^{-1}(W^\top)$ must be identical.

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_{ts}—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_{ts} 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 $k' \leq k$ and $i' \leq i$. 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 $L \geq k \geq j \geq i \geq 0$, the dimensions α_{kji} of the flow subspaces A_{kji}, the dimensions β_{kji} of the flow subspaces B_{kji}, and the dimensions ω_{ts} of the prebasis subspaces a_{tss} and b_{tss} are related by the identities*

$$\alpha_{kji} = \dim A_{kji} = \sum_{t=j}^{k} \sum_{s=0}^{i} \omega_{ts} \quad \text{and} \quad \beta_{kji} = \dim B_{kji} = \sum_{t=k}^{L} \sum_{s=i}^{j} \omega_{ts}. \tag{4.8}$$

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 (4.8) into the right-hand side, we obtain (4.8) 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 (4.8) 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 $\mathbb{R}^{d_j}$. It implies that every valid rank list induces a valid multiset of intervals (as we defined a multiset of intervals to be *valid* if it satisfies the following identity).
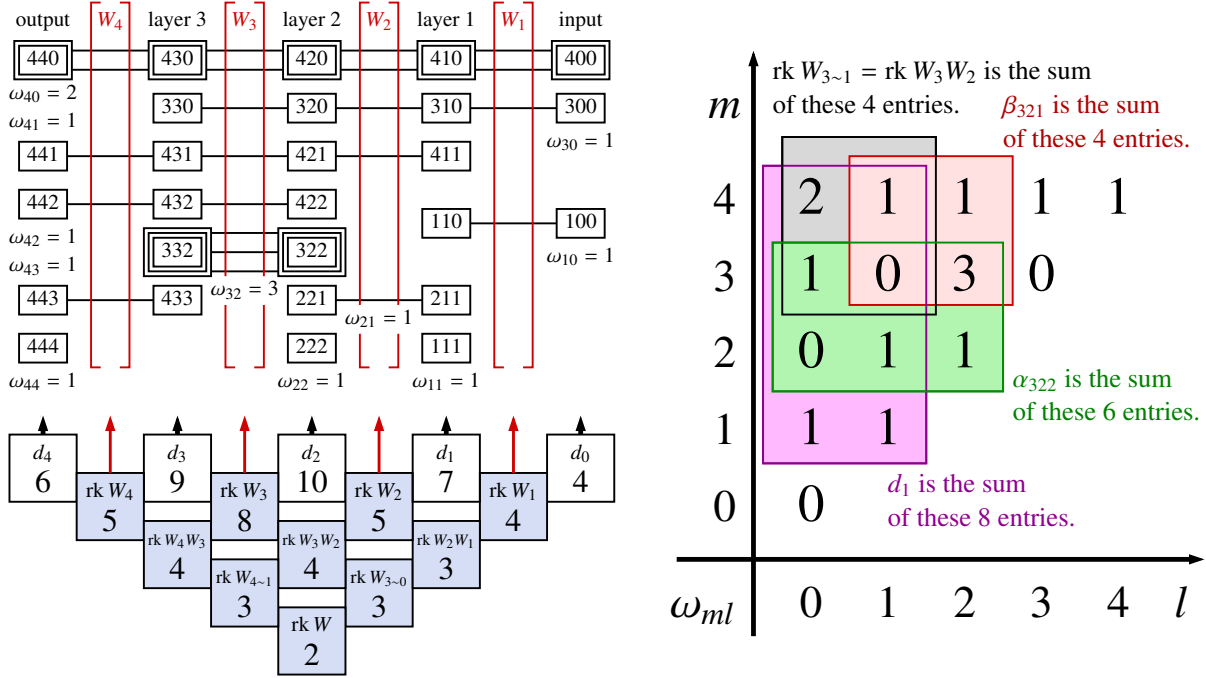
$$d_j = \sum_{t=j}^{L} \sum_{s=0}^{j} \omega_{ts} \quad \bigg| \quad \text{rk } W_{k\sim i} = \sum_{t=k}^{L} \sum_{s=0}^{i} \omega_{ts} \quad \bigg| \quad \alpha_{kji} = \sum_{t=j}^{k} \sum_{s=0}^{i} \omega_{ts} \quad \bigg| \quad \beta_{kji} = \sum_{t=k}^{L} \sum_{s=i}^{j} \omega_{ts}$$

Figure 7: At left, we reprise the basis flow diagram from Figure 5. At right, we tabulate the values of the interval multiplicities ω_{ts} with boxes that illustrate how the four summations compute d_1, α_{322}, β_{321}, and rk $W_{3\sim1}$ = rk $W_3 W_2$. At the bottom of the figure, we reprise the four summations for reference.

Corollary 9. *The number of units in unit layer j is, as formula (4.1) says,*

$$d_j = \sum_{t=j}^{L} \sum_{s=0}^{j} \omega_{ts}.$$

Proof. As $\mathbb{R}^{d_j} = A_{Ljj} = B_{jj0}$, $d_j = \alpha_{Ljj} = \beta_{jj0}$. The summation follows by identity (4.8). $\qed$

Recall that a rank list $\underline{r} = \langle \text{rk } W_{k\sim i} \rangle_{L \geq k \geq i \geq 0}$ is a list of the ranks of all the subsequence matrices, including those of the form rk $W_{j\sim j} = d_j$. 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 (4.9) and (4.10).

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*

$$\text{rk } W_{k\sim i} \quad = \quad \alpha_{Lki} = \beta_{ki0} = \sum_{t=k}^{L} \sum_{s=0}^{i} \omega_{ts} \quad \text{and} \tag{4.9}$$

$$\omega_{ki} \quad = \quad \text{rk } W_{k\sim i} - \text{rk } W_{k\sim i-1} - \text{rk } W_{k+1\sim i} + \text{rk } W_{k+1\sim i-1}, \tag{4.10}$$

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

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 (4.8) to connect those to the interval multiplicities. Recall that $A_{Lii} = \mathbb{R}^{d_i}$ and $A_{k-1,i,i} = \text{null } W_{k \sim i} \cap \text{col } W_{i \sim i} = \text{null } W_{k \sim i}$. As $W_{k \sim i}$ is a $d_k \times d_i$ matrix,

$$
\begin{aligned}
\text{rk } W_{k \sim i} &= d_i - \dim \text{null } W_{k \sim i} \\
&= \dim A_{Lii} - \dim A_{k-1,i,i} \\
&= \alpha_{Lii} - \alpha_{k-1,i,i} \\
&= \sum_{t=k}^{L} \sum_{s=0}^{i} \omega_{ts} \\
&= \alpha_{Lki} = \beta_{ki0}
\end{aligned}
$$

as claimed. (Symmetrically, we could obtain the summation (4.9) by instead starting from $\text{rk } W_{k \sim i} = d_k - \dim \text{null } W_{k \sim i}^{\top}$ and recalling that $B_{kk0} = \mathbb{R}^{d_k}$ and $B_{k,k,i+1} = \text{null } W_{k \sim i}^{\top}$. This is how we originally realized that $\dim a_{kji} = \dim b_{kji}$, which led us to Lemma 2.)

We can verify the identity $\text{rk } W_{k \sim i} - \text{rk } W_{k \sim i-1} - \text{rk } W_{k+1 \sim i} + \text{rk } W_{k+1 \sim i-1} = \omega_{ki}$ by substituting the summation (4.9) into it. $\qquad \square$

Ferdinand Georg Frobenius [6] proved in 1911 that $\text{rk } W_{k \sim i} - \text{rk } W_{k \sim i-1} - \text{rk } W_{k+1 \sim i} + \text{rk } W_{k+1 \sim i-1} \geq 0$, a statement called the *Frobenius rank inequality*. This confirms that every ω_{ts} is nonnegative (a fact we already knew, as every prebasis subspace has a nonnegative dimension). Our derivations deepen the Frobenius rank inequality by connecting the slack (4.10) in the inequality to the dimension of the subspaces (4.6) and (4.7). (We considered calling each ω_{ts} a *Frobenius slack* instead of an *interval multiplicity*.) To put it in a simpler notation, for any four matrices R, S, T, and U,

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

Thome [20] offers some generalizations of the Frobenius rank inequality to linear neural networks. He proves them by induction, but they also follow easily from (4.9).

4.6 The Canonical Weight Vector

Figure 8 depicts a sequence of matrices we call *almost-identity matrices*, which we define to be any matrix obtained by taking an identity matrix, inserting additional rows of zeros, and appending additional columns of zeros on the right. Products of almost-identity matrices are also almost-identity matrices. For any given valid rank list $\underline{r}$, we will define one unique *canonical weight vector*

$$
\tilde{\theta} = (\tilde{I}_L, \tilde{I}_{L-1}, \ldots, \tilde{I}_1)
$$

whose rank list is $\underline{r}$ and whose matrices $\tilde{I}_j$ are almost-identity matrices, as illustrated in Figure 8. It is "canonical" in the sense that it depends solely on $\underline{r}$. Let $\tilde{I} = \mu(\tilde{\theta}) = \tilde{I}_L \tilde{I}_{L-1} \cdots \tilde{I}_1$. Our canonical weight vectors will have the property that the sole nonzero components of $\tilde{I}$ are on its diagonal, in the uppermost (and leftmost) positions on the diagonal.

This section is devoted to showing that for any weight vector θ with rank list $\underline{r}$, there is a straightforward relationship between θ, the canonical weight vector $\tilde{\theta}$ for the rank list $\underline{r}$, and any set of flow prebases $\mathcal{A}_j$, $j \in [0, L]$ associated with θ. Specifically, the flow prebases induce a linear transformation that maps θ to $\tilde{\theta}$.
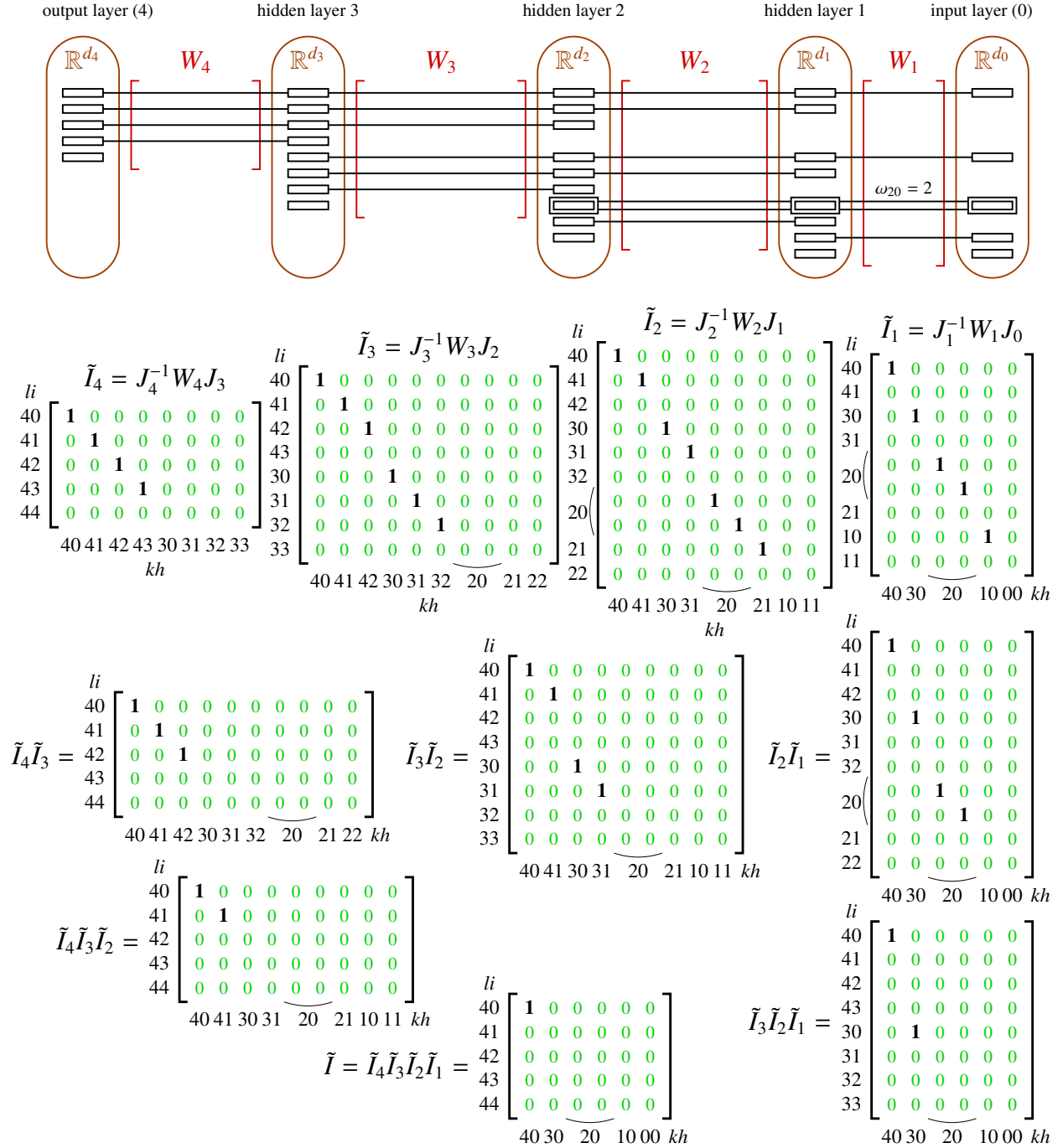
Figure 8: The canonical weight vector $\tilde{\theta} = (\tilde{I}_4, \tilde{I}_3, \tilde{I}_2, \tilde{I}_1)$ and its subsequence matrices when every interval multiplicity is $\omega_{ki} = 1$ except $\omega_{20} = 2$. Every rank-1 matrix $W \in \mathbb{R}^{5 \times 6}$ can be factored as $W = J_4 \tilde{I}_4 \tilde{I}_3 \tilde{I}_2 \tilde{I}_1 J_0^{-1}$ where the matrices $\tilde{I}_j$ have the values depicted (determined solely by the choice of interval multiplicities) but J_4 and J_0 depend on W.

The immediate benefit of this relationship is that it reveals a way to convert flow prebases $\mathcal{A}_j$ to flow prebases $\mathcal{B}_j$ for the transpose network, or vice versa (see Section 4.7). More importantly, in Section 5.2 we will use the canonical weight vector to show that any two strata (from different fibers) with the same rank list are related by an invertible linear transformation. Hence, the topology of a stratum depends solely on its rank list. The canonical weight vector also gives us intuition about the geometry of each stratum (see Section 7.7). Lastly, for any valid multiset of intervals, the canonical weight vector gives an explicit example of a weight vector having those intervals (and the corresponding rank list).

Consider a weight vector $\theta = (W_L, W_{L-1}, \ldots, W_1) \in \mathbb{R}^{d_\theta}$. Suppose we choose flow prebases $\mathcal{A}_0, \mathcal{A}_1, \ldots, \mathcal{A}_L$ (which depend in part on θ) as described in Section 4.4; Lemma 7 guarantees we can. We chose the name "prebasis" because we can convert the prebasis $\mathcal{A}_j$ into a basis for $\mathbb{R}^{d_j}$, and now we will use that basis, in the form of a square, invertible matrix J_j whose columns are the basis vectors. Recall that $\mathcal{A}_j$ is a prebasis for $\mathbb{R}^{d_j}$ (by Lemma 4, as $\mathcal{A}_j = \mathcal{A}_{Ljj}$ and $A_{Ljj} = \mathbb{R}^{d_j}$), so $\mathbb{R}^{d_j} = \bigoplus_{k=j}^{L} \bigoplus_{i=0}^{j} a_{kji}$, the vector sum of the subspaces in $\mathcal{A}_j$. Forming J_j is a matter of picking basis vectors for the subspaces a_{kji}—but we want those basis vectors to *flow*, just like the subspaces do. As we have chosen flow prebases, the prebasis subspaces satisfy $a_{kji} = W_{j\sim i} a_{kii}$. For each subspace $a_{kii} \in \mathcal{A}_i$ with nonzero dimension ω_{ki}, choose an arbitrary set of ω_{ki} vectors in $\mathbb{R}^{d_i}$ that form a basis for a_{kii}, then let J_{kii} be a $d_i \times \omega_{ki}$ matrix whose columns are those basis vectors. To obtain *flow bases*, for each subspace $a_{kji} \in \mathcal{A}_j$ with $j > i$ and nonzero dimension ω_{ki}, we set

$$J_{kji} = W_{j\sim i} J_{kii},$$

so J_{kji} is a $d_j \times \omega_{ki}$ matrix whose columns are a basis for a_{kji}. We construct a $d_j \times d_j$ matrix J_j whose columns are a basis for $\mathbb{R}^{d_j}$ by grouping together all the matrices J_{kji} with matching j. For example, if $L = 4$ then

$$J_2 = [J_{420} \quad J_{421} \quad J_{422} \quad J_{320} \quad J_{321} \quad J_{322} \quad J_{220} \quad J_{221} \quad J_{222}].$$

Note that some of the blocks in this matrix may be empty; for instance, if $\omega_{31} = 0$, then J_{321} contributes nothing to J_2. In general, we order the matrices J_{kji} primarily in order of decreasing k and secondarily in order of increasing i, because it gives us the almost-identity matrix structure depicted in Figure 8. As $\mathcal{A}_j$ is a prebasis for $\mathbb{R}^{d_j}$, the a_{kji}'s are linearly independent (for a fixed j and varying k, i), so the columns of J_j are linearly independent. Hence the columns of J_j are a basis for $\mathbb{R}^{d_j}$ and J_j is invertible.

We now define the *canonical weight vector* to be $\tilde{\theta} = (\tilde{I}_L, \tilde{I}_{L-1}, \ldots, \tilde{I}_1)$, where

$$\tilde{I}_j = J_j^{-1} W_j J_{j-1}.$$

With these, we can factor $W = \mu(\theta)$ into $L + 2$ matrices of the form

$$W = J_L \tilde{I}_L \tilde{I}_{L-1} \cdots \tilde{I}_1 J_0^{-1} \tag{4.11}$$

because $J_L \tilde{I}_L \tilde{I}_{L-1} \cdots \tilde{I}_1 J_0^{-1} = J_L J_L^{-1} W_L J_{L-1} J_{L-1}^{-1} W_{L-1} J_{L-2} \cdots J_1 W_1 J_0 J_0^{-1} = W_L W_{L-1} \cdots W_1 = W$. This factorization includes two matrices that are not canonical, J_L and J_0^{-1} (which depend on the value of W and on some arbitrary choices of prebasis subspaces and basis vectors).

We now show that the matrices $\tilde{I}_j$ are almost-identity matrices and they are *canonical*: they depend solely on our choice of interval multiplicities (equivalently, on the rank list of θ), and they are independent of W except for the fact that $\omega_{L0} = \text{rk } W$ is fixed. To understand the almost-identity matrices, we number their rows and columns in a non-standard way, illustrated in Figure 8. For each interval $[i, l]$ such that $L \geq l \geq j \geq i \geq 0$, the matrix $\tilde{I}_j$ has ω_{li} rows representing that interval; these rows have index li in Figure 8. We order the rows of $\tilde{I}_j$ primarily in order of decreasing l and secondarily in order of increasing i (as illustrated, to match the

ordering in J_j). Similarly, for each interval $[h, k]$ such that $L \geq k \geq j - 1 \geq h \geq 0$, $\tilde{I}_j$ has ω_{kh} columns representing that interval; these columns have index kh in the figure. We order these columns primarily in order of decreasing k and secondarily in order of increasing h (as illustrated, to match the ordering in J_{j-1}). The following lemma describes the structure of $\tilde{I}_j$, as illustrated in Figure 8.

Lemma 11. *For every interval $[i, k]$ that contains both $j - 1$ and j and has multiplicity $\omega_{ki} > 0$, $\tilde{I}_j$ has an $\omega_{ki} \times \omega_{ki}$ block that is an identity matrix, located at the rows and columns associated with the interval $[i, k]$. All the other components of $\tilde{I}_j$ (where the column and the row do not represent the same interval) are zero.*

Proof. We prove the lemma by showing that if we construct $\tilde{I}_j$ as described by the lemma, then $J_j \tilde{I}_j = W_j J_{j-1}$. This confirms that $\tilde{I}_j = J_j^{-1} W_j J_{j-1}$ has the structure we claim it has.

Let $x \in \mathbb{R}^{d_{j-1}}$ be a unit coordinate vector: a vector whose components are all 0's except that one component is a 1. Then $J_{j-1} x$ is a column of J_{j-1} (and one of our selected basis vectors for $\mathbb{R}^{d_{j-1}}$). Moreover, it is a column of some matrix $J_{k,j-1,i}$ that is a block in J_{j-1}; suppose $J_{j-1} x$ is the zth column of $J_{k,j-1,i}$. With respect to the columns of the almost-identity matrix $\tilde{I}_j$ as described above, the 1 component in x is aligned with the zth column among the columns that represent the interval $[i, k]$.

Consider the product $\tilde{I}_j x$ in two cases. If $k \geq j$ then $\tilde{I}_j x$ is a unit coordinate vector: with respect to the rows of $\tilde{I}_j$ as described above, the 1 component in $\tilde{I}_j x$ is aligned with the zth row among the rows of $\tilde{I}_j$ that represent the interval $[i, k]$. Hence $J_j \tilde{I}_j x$ is the zth column of J_{kji}. Recall that $J_{kji} = W_j J_{k,j-1,i}$; it follows that $J_j \tilde{I}_j x = W_j J_{j-1} x$. The only other case is that $k = j - 1$; then $\tilde{I}_j x = \mathbf{0}$ because the interval $[i, k]$ is not represented among the rows of $\tilde{I}_j$. In that case, $W_j J_{j-1} x = \mathbf{0}$ because $J_{j-1} x \in a_{j-1,j-1,i}$ and $W_j a_{j-1,j-1,i} = \{\mathbf{0}\}$; so again, $J_j \tilde{I}_j x = W_j J_{j-1} x$.

The fact that $J_j \tilde{I}_j x = W_j J_{j-1} x$ for every unit coordinate vector $x \in \mathbb{R}^{d_{j-1}}$ implies that $J_j \tilde{I}_j = W_j J_{j-1}$. $\square$

4.7   The Transpose Network and Complementary Flow Bases

Section 4.6 describes a relationship between a weight vector θ, the canonical weight vector $\tilde{\theta}$ with the same rank list, and a set of flow prebases $\mathcal{A}_j$ associated with θ. Can we use flow prebases $\mathcal{B}_j$ associated with the transpose network instead? We can, and doing so reveals a connection between the two sets of flow prebases.

Echoing the construction in Section 4.6, we build a basis for each subspace b_{kji} by choosing ω_{ki} basis vectors for b_{kji} and writing them as the columns of a $d_j \times \omega_{ki}$ matrix K_{kji} chosen to satisfy the flow condition for the transpose network,

$$K_{kji} = W_{k \sim j}^{\top} K_{kki}.$$

Then we append them together to build a basis K_j for each $\mathbb{R}^{d_j}$. For example, if $L = 4$ then

$$K_2 = [K_{420} \quad K_{421} \quad K_{422} \quad K_{320} \quad K_{321} \quad K_{322} \quad K_{220} \quad K_{221} \quad K_{222}].$$

By a straightforward reflection of the arguments given in Section 4.6, we can show that by writing

$$\tilde{I}_j = K_j^{\top} W_j K_{j-1}^{-\top},$$

we obtain the same canonical almost-identity matrices $\tilde{I}_j$ as we did in Section 4.6. Hence we can factor $W = \mu(\theta)$ into $L + 2$ matrices of the form

$$W = K_L^{-\top} \tilde{I}_L \tilde{I}_{L-1} \cdots \tilde{I}_1 K_0^{\top}.$$

There are (usually) many possible choices of flow prebases b_{kji}, and there are many possible choices of flow bases K_j. But they all produce the same canonical matrices $\tilde{I}_j$. Conversely, any set of invertible matrices K_j that produce the canonical matrices $\tilde{I}_j$ defines a set of flow bases. Recall from Section 4.6 that one way to produce such an $\tilde{I}_j$ is to write $\tilde{I}_j = J_j^{-1} W_j J_{j-1}$. It follows that if you have flow bases $J_0, J_1, \ldots, J_L$ for the forward network, then you can find flow bases for the transpose network by choosing

$$K_j = J_j^{-\top}. \tag{4.12}$$

Symmetrically, if you have flow bases $K_0, K_1, \ldots, K_L$ for the transpose network, then $J_j = K_j^{-\top}$ are flow bases for the forward network.

With that choice, $J_j^\top K_j = I$, hence

$$J_{kji}^\top K_{k'ji'} = \begin{cases} I, & k = k' \text{ and } i = i', \\ 0, & \text{otherwise.} \end{cases}$$

Therefore, (for this choice of flow prebases) if $k \neq k'$ or $i \neq i'$, every vector in the subspace a_{kji} is orthogonal to every vector in $b_{k'ji'}$. Hence if you have flow prebases $\mathcal{A}_0, \mathcal{A}_1, \ldots, \mathcal{A}_L$ for the forward network, then you can find flow prebases $\mathcal{B}_0, \mathcal{B}_1, \ldots, \mathcal{B}_L$ for the transpose network by choosing each b_{kji} to be the unique subspace of dimension ω_{ki} that satisfies these orthogonality properties; that is,

$$b_{kji} = \left(\sum_{k' \neq k \text{ or } i' \neq i} a_{k'ji'} \right)^\perp. \tag{4.13}$$

(Note that this expression does not depend on the specific bases J_j chosen to express the prebases $\mathcal{A}_j$.) This transformation from $\mathcal{A}_j$ to $\mathcal{B}_j$ serves as its own inverse, converting flow prebases for the transpose network to flow prebases for the forward network.

We say that the flow prebases $\mathcal{B}_0, \mathcal{B}_1, \ldots, \mathcal{B}_L$ generated by (4.13)—or, equivalently, by (4.12)—are *complementary* to the flow prebases $\mathcal{A}_0, \mathcal{A}_1, \ldots, \mathcal{A}_L$. We say that the flow bases $K_0, K_1, \ldots, K_L$ generated by (4.12) are *complementary* to the flow bases $J_0, J_1, \ldots, J_L$. Note that complementary flow bases imply complementary flow prebases, but not vice versa.

4.8   A Fundamental Theorem of Linear Neural Networks?

Matrices have crucial properties that Gilbert Strang [18] summarizes as a *Fundamental Theorem of Linear Algebra*, describing the relationships between the four fundamental subspaces of a matrix: the rowspace, the columnspace, the nullspace, and the left nullspace. For any $p \times q$ matrix W, the rowspace of W is the orthogonal complement of the nullspace of W and the sum of their dimensions is q (the latter fact is known as the *Rank-Nullity Theorem*). The same observations apply to $W^\top$, so the columnspace of W is the orthogonal complement of the left nullspace of W and the sum of their dimensions is p. The dimensions of row W and col W are the same, and we call that dimension rk W. (Strang's "Fundamental Theorem" also addresses properties of the singular value decomposition, not treated here.)

Here, we outline a candidate for an analogous *Fundamental Theorem of Linear Neural Networks*. Theorem 12 below takes parts of Lemmas 3, 4, 6, and 7, and the formulae (4.1), (4.2), (4.8), (4.9), and (4.10), and reframes them around the decomposition of each layer of units into "fundamental" flow subspaces akin to row W, col W, null W, and null $W^\top$ (though our decomposition into flow subspaces is not generally unique).

Theorem 12. *Consider a weight vector $\theta = (W_L, W_{L-1}, \ldots, W_1)$ representing a linear neural network. For $L \geq k \geq i \geq 0$, let*

$$\omega_{ki} = \text{rk } W_{k \sim i} - \text{rk } W_{k \sim i-1} - \text{rk } W_{k+1 \sim i} + \text{rk } W_{k+1 \sim i-1}.$$

Then there exist for $L \geq k \geq j \geq i \geq 0$ subspaces a_{kji} and b_{kji} of dimension ω_{ki} such that for all $j \in [0, L]$, $\mathcal{A}_j = \{a_{kji} \neq \{\mathbf{0}\} : k \in [j, L], i \in [0, j]\}$ is a direct sum decomposition of $\mathbb{R}^{d_j}$, $\mathcal{B}_j = \{b_{kji} \neq \{\mathbf{0}\} : k \in [j, L], i \in [0, j]\}$ is also a direct sum decomposition of $\mathbb{R}^{d_j}$, and the subspaces satisfy the flow *conditions*

$$W_j a_{k,j-1,i} \;\; = \;\; \begin{cases} a_{kji}, & k \geq j, \\ \{\mathbf{0}\}, & k = j - 1, \end{cases} \qquad and$$

$$W_j^\top b_{kji} \;\; = \;\; \begin{cases} b_{k,j-1,i}, & j > i, \\ \{\mathbf{0}\}, & j = i. \end{cases}$$

Moreover,

$$A_{kji} = \text{null } W_{k+1 \sim j} \cap \text{col } W_{j \sim i} \;\; = \;\; \bigoplus_{t=j}^{k} \bigoplus_{s=0}^{i} a_{tjs}, \quad k \in [j-1, L], i \in [-1, j], \quad and$$

$$B_{kji} = \text{row } W_{k \sim j} \cap \text{null } W_{j \sim i-1}^\top \;\; = \;\; \bigoplus_{t=k}^{L} \bigoplus_{s=i}^{j} b_{tjs}, \quad k \in [j, L+1], i \in [0, j+1].$$

The dimensions of A_{kji} and B_{kji} are, respectively, $\alpha_{kji} = \sum_{t=j}^{k} \sum_{l=0}^{i} \omega_{ts}$ and $\beta_{kji} = \sum_{t=k}^{L} \sum_{s=i}^{j} \omega_{ts}$. Moreover, $d_j = \sum_{t=j}^{L} \sum_{s=0}^{j} \omega_{ts}$ and $\text{rk } W_{k \sim i} = \sum_{t=k}^{L} \sum_{s=0}^{i} \omega_{ts}$. Note that this decomposition is not necessarily unique, but the subspace dimensions ω_{ki}, α_{kji}, and β_{kji} are the same for all such decompositions.

Moreover, given prebases $\mathcal{A}_j$ that satisfy the conditions above, we can obtain prebases $\mathcal{B}_j$ that satisfy them by choosing

$$b_{kji} = \left(\sum_{(k', i') \neq (k, i)} a_{k' j i'} \right)^\perp.$$

We say these $\mathcal{B}$-prebases are complementary *to the $\mathcal{A}$-prebases. By a symmetric formula, given prebases $\mathcal{B}_j$ that satisfy the conditions above, we can obtain (complementary) prebases $\mathcal{A}_j$ that satisfy them.*

5 Strata in the Rank Stratification

In this section, we prove some of the results about rank stratifications promised in Section 3.

- Each stratum $S_{\underline{r}}$—the set of points on a fiber $\mu^{-1}(W)$ with rank list $\underline{r}$—is a smooth manifold (Theorem 16).

- Given two strata from two different fibers that have the same rank list, there is a linear homeomorphism mapping one stratum to the other (Corollary 14). Hence the topology of a stratum is solely determined by its rank list.

- Given two points on the same stratum, there is a linear homeomorphism mapping the stratum to itself and mapping one point to the other (Corollary 15).

5.1 Strata, Semi-Algebraic Sets, and Determinantal Varieties

Let $\mathcal{S}$ be a rank stratification of a fiber $\mu^{-1}(W)$. Here we will see that although a stratum in $\mathcal{S}$ is not necessarily an algebraic variety, it belongs to the closely related class of semi-algebraic sets. This will help us to see in Section 5.2 why each stratum in $\mathcal{S}$ is a manifold. A *semi-algebraic set* is a set that can be obtained from finitely many sets of the form $\{\theta \in \mathbb{R}^{d_\theta} : f_i(\theta) \geq 0\}$, where each f_i is a polynomial, by a combination of union, intersection, and complementation operations. This class of sets include all algebraic varieties, which can be expressed as intersections of sets of the form $\{\theta \in \mathbb{R}^{d_\theta} : f_i(\theta) = 0\}$, each of which is the intersection of two sets of the form $\{\theta \in \mathbb{R}^{d_\theta} : f_i(\theta) \geq 0\}$.

To see that a stratum is an algebraic set, we must first see how to constrain the rank of a matrix with algebra. The constraint that a matrix M has rank at most r can be written as a constraint that the determinant of every $(r + 1) \times (r + 1)$ minor of M is zero. (The number of polynomial equations specified this way can grow exponentially with r, but typically most of them are redundant.) Thus the set of all $p \times q$ matrices with rank r or less is an algebraic variety, called the *determinantal variety* [9, Lecture 9], which we denote

$$\mathrm{DV}_r^{p \times q} = \{M \in \mathbb{R}^{p \times q} : \mathrm{rk}\, M \leq r\}.$$

The determinantal variety has singular points and thus is not a manifold (unless r is zero; $\mathrm{DV}_0^{p \times q}$ contains only the zero matrix). It is well known that the singular locus of $\mathrm{DV}_r^{p \times q}$ is $\mathrm{DV}_{r-1}^{p \times q}$—the matrices with rank strictly less than r. If we omit those matrices, we obtain a manifold that we call the *determinantal manifold*,

$$\mathrm{DM}_r^{p \times q} = \mathrm{DV}_r^{p \times q} \setminus \mathrm{DV}_{r-1}^{p \times q} = \{M \in \mathbb{R}^{p \times q} : \mathrm{rk}\, M = r\}.$$

It is not hard to see that $\mathrm{DV}_r^{p \times q}$ is the closure of $\mathrm{DM}_r^{p \times q}$.[5] So although $\mathrm{DM}_r^{p \times q}$ is a manifold, it is not a closed point set with respect to the weight space (unless $r = 0$). It is well known that $\mathrm{DV}_r^{p \times q}$ has dimension $r(p + q - r)$, hence so does $\mathrm{DM}_r^{p \times q}$.

$\mathrm{DM}_r^{p \times q}$ is not an algebraic variety (unless $r = 0$). A $p \times q$ matrix M is in $\mathrm{DM}_r^{p \times q}$ if and only if the determinant of every $(r + 1) \times (r + 1)$ minor of M is zero but the determinant of at least one $r \times r$ minor is nonzero. We cannot express the latter constraint in a system of polynomial equations. But $\mathrm{DM}_r^{p \times q}$ is a set difference of two algebraic varieties, so it is a semi-algebraic set.

We define a stratum in $\mathcal{S}$ by fixing the rank of each subsequence matrix. If a weight vector $\theta = (W_L, \ldots, W_1)$ has rank list $\underline{r}$, then W_1 lies on the determinantal manifold $\mathrm{DM}_{r_{1 \sim 0}}$, $W_2 W_1$ lies on $\mathrm{DM}_{r_{2 \sim 0}}$, $W_4 W_3 W_2$ lies on $\mathrm{DM}_{r_{4 \sim 1}}$, and so on. These constraints are semi-algebraic, so a stratum is a semi-algebraic set.

For our purposes, we need to express these constraints in terms of the weight vector θ, not in terms of $W_2 W_1$ and $W_4 W_3 W_2$. This motivates what we call a *weight-space determinantal manifold*, denoted

$$\mathrm{WDM}_{\underline{r}}^{k \sim i} = \{(W_L, \ldots, W_1) : \mathrm{rk}\, W_{k \sim i} = r_{k \sim i}\}.$$

One difference between $\mathrm{WDM}_{\underline{r}}^{k \sim i}$ and $\mathrm{DM}_r^{d_k \times d_i}$ is that $\mathrm{WDM}_{\underline{r}}^{k \sim i}$ is a set of points in weight space, and $\mathrm{DM}_r^{d_k \times d_i}$ is a set of points in matrix space. The subsequence matrices are polynomial in θ, so $\mathrm{WDM}_{\underline{r}}^{k \sim i}$ is a semi-algebraic set that can be expressed as a set difference of two algebraic varieties, just like $\mathrm{DM}_r^{p \times q}$.

Consider partitioning the entire weight space $\mathbb{R}^{d_\theta}$ by rank list (forgetting briefly about the fiber and the strata). Consider the set of all points in weight space having a specified rank list $\underline{r}$. We call this set a *multideterminantal manifold*, denoted

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

---

[5]The set $\{\mathrm{DM}_0^{p \times q}, \mathrm{DM}_1^{p \times q}, \ldots, \mathrm{DM}_r^{p \times q}\}$ is a stratification of $\mathrm{DV}_r^{p \times q}$ into manifolds of dimension $0, 1, \ldots, r$.

The multideterminantal manifold is the intersection of the weight-space determinantal manifolds:

$$\text{MDM}_{\underline{r}} = \bigcap_{L \geq k > i \geq 0} \text{WDM}_{\underline{r}}^{k \sim i}.$$

It follows that $\text{MDM}_{\underline{r}}$ is a semi-algebraic set. ($\text{MDM}_{\underline{r}}$ is also a manifold, but we will not prove it, as we don't need to. But in Section 5.2 we prove that each stratum is a manifold, and that proof can easily be adapted to $\text{MDM}_{\underline{r}}$.) The entire weight space $\mathbb{R}^{d_\theta}$ can be partitioned into a finite set of these multideterminantal manifolds, one for each valid rank list.

Each stratum in the rank stratification $\mathcal{S}$ is the intersection of the fiber with a multideterminantal manifold,

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

Hence, $S_{\underline{r}}^{W}$ is a semi-algebraic set. We prove in the next section that it is a manifold.

5.2 The Rank List Solely Determines a Stratum's Topology

Here we show that in the rank stratification, each stratum is a smooth manifold, and moreover, that the topology of a stratum is determined solely by its rank list. In particular, if two strata in two different fibers have the same rank list, a linear transformation maps one stratum to the other. This linear transformation (restricted to the first stratum) is a homeomorphism, showing that the topology of a stratum is fully determined by its rank list. Moreover, this transformation can be chosen to map any selected point on one stratum to any selected point on the other. Therefore, the local topology of a stratum looks the same from every point on the stratum, and even the geometry of a stratum looks the same from every point if we ignore scaling. A semi-algebraic set with this property is necessarily a smooth manifold.

Consider a matrix W and a point on its fiber, $\theta = (W_L, W_{L-1}, \ldots, W_1) \in \mu^{-1}(W)$, with rank list $\underline{r}$. In the rank stratification $\mathcal{S}$ of $\mu^{-1}(W)$, $S_{\underline{r}}^{W}$ is the stratum that contains θ. Define the basis matrices $J_L, J_{L-1}, \ldots, J_0$ described in Section 4.6, noting that these matrices depend in part on our choice of θ. Recall from Section 4.6 the almost-identity matrices $\tilde{I}_j = J_j^{-1} W_j J_{j-1}$ and the canonical weight vector

$$\tilde{\theta} = (\tilde{I}_L, \tilde{I}_{L-1}, \ldots, \tilde{I}_1).$$

Recall that the value of $\tilde{\theta}$ depends solely on $\underline{r}$; it does not otherwise depend on θ nor the product matrix W.

Consider the linear transformations

$$\eta : \mathbb{R}^{d_\theta} \to \mathbb{R}^{d_\theta}, (M_L, M_{L-1}, \ldots, M_1) \mapsto (J_L M_L J_{L-1}^{-1}, J_{L-1} M_{L-1} J_{L-2}^{-1}, \ldots, J_1 M_1 J_0^{-1}) \quad \text{and} \quad (5.1)$$

$$\eta^{-1} : \mathbb{R}^{d_\theta} \to \mathbb{R}^{d_\theta}, (M_L, M_{L-1}, \ldots, M_1) \mapsto (J_L^{-1} M_L J_{L-1}, J_{L-1}^{-1} M_{L-1} J_{L-2}, \ldots, J_1^{-1} M_1 J_0).$$

Observe that $\eta^{-1}(\theta) = \tilde{\theta}$ and $\eta(\tilde{\theta}) = \theta$. As the J_i's are invertible, η and η^{-1} are linear bijections from weight space to weight space. Crucially, it is easy to see that ϕ, $\eta(\phi)$, and $\eta^{-1}(\phi)$ all have the same rank list for any $\phi \in \mathbb{R}^{d_\theta}$, as a matrix's rank is invariant under multiplication by an invertible matrix.

Let $\tilde{I} = \mu(\tilde{\theta}) = \tilde{I}_L \tilde{I}_{L-1} \cdots \tilde{I}_1$, depicted in Figure 8. Now consider the fiber $\mu^{-1}(\tilde{I})$ of $\tilde{I}$ and some stratum $S_{\underline{s}}^{\tilde{I}}$ in the rank stratification of that fiber. The following lemma shows that the linear transformation η maps the fiber $\mu^{-1}(\tilde{I})$ to the fiber $\mu^{-1}(W)$, and likewise maps the stratum $S_{\underline{s}}^{\tilde{I}}$ to the stratum $S_{\underline{s}}^{W}$. So both fibers have the same topology, and both strata have the same topology. Given a set $Z \subseteq \mathbb{R}^{d_\theta}$, let $\eta(Z)$ denote the set obtained by applying η to every weight vector in Z.

Lemma 13. $\eta(\mu^{-1}(\tilde{I})) = \mu^{-1}(W)$ and $\eta\left(S_{\underline{s}}^{\tilde{I}}\right) = S_{\underline{s}}^{W}$ for every stratum $S_{\underline{s}}^{\tilde{I}}$ in the rank stratification of $\mu^{-1}(\tilde{I})$.

Proof. For any point $\phi = (X_L, X_{L-1}, \ldots, X_1) \in \mathbb{R}^{d_\theta}$,

$$\mu(\eta(\phi)) = J_L X_L J_{L-1}^{-1} J_{L-1} X_{L-1} J_{L-2}^{-1} \cdots J_1 X_1 J_0^{-1} = J_L X_L X_{L-1} \cdots X_1 J_0^{-1} = J_L \mu(\phi) J_0^{-1} \quad \text{and}$$

$$\mu(\eta^{-1}(\phi)) = J_L^{-1} X_L J_{L-1} J_{L-1}^{-1} X_{L-1} J_{L-2} \cdots J_1^{-1} X_1 J_0 = J_L^{-1} X_L X_{L-1} \cdots X_1 J_0 = J_L^{-1} \mu(\phi) J_0.$$

From (4.11), we have $W = J_L \tilde{I} J_0^{-1}$, and thus $\tilde{I} = J_L^{-1} W J_0$. To see that η maps any point on $\mu^{-1}(\tilde{I})$ to a point on $\mu^{-1}(W)$, observe that for any point $\phi \in \mu^{-1}(\tilde{I})$, $\mu(\eta(\phi)) = J_L \mu(\phi) J_0^{-1} = J_L \tilde{I} J_0^{-1} = W$. Conversely, to see that η^{-1} maps any point on $\mu^{-1}(W)$ to a point on $\mu^{-1}(\tilde{I})$, observe that for any point $\phi \in \mu^{-1}(W)$, $\mu(\eta^{-1}(\phi)) = J_L^{-1} \mu(\phi) J_0 = J_L^{-1} W J_0 = \tilde{I}$. Hence, $\eta(\mu^{-1}(\tilde{I})) = \mu^{-1}(W)$ as claimed.

Recall that $S_{\underline{s}}^W$ contains the points on $\mu^{-1}(W)$ with rank list $\underline{s}$ and $S_{\underline{s}}^{\tilde{I}}$ contains the points on $\mu^{-1}(\tilde{I})$ with rank list $\underline{s}$. As ϕ, $\eta(\phi)$, and $\eta^{-1}(\phi)$ have the same rank list for any $\phi \in \mathbb{R}^{d_\theta}$, η maps any point on $S_{\underline{s}}^{\tilde{I}}$ to a point on $S_{\underline{s}}^W$, η^{-1} maps any point on $S_{\underline{s}}^W$ to a point on $S_{\underline{s}}^{\tilde{I}}$, and $\eta\left(S_{\underline{s}}^{\tilde{I}}\right) = S_{\underline{s}}^W$ as claimed. $\square$

See Appendix E for another lemma related to Lemma 13, applied to a subspace we will introduce in Section 8.1.

It is a short extra step to see that if we replace $\tilde{I}$ with any other matrix W' of the same rank, there exists a linear transformation that likewise maps $\mu^{-1}(W)$ to $\mu^{-1}(W')$ and $S_{\underline{s}}^W$ to $S_{\underline{s}}^{W'}$. Therefore all strata with the same rank list have essentially the same geometry, up to a linear transformation in weight space.

Corollary 14. *Consider two matrices W and W' (not necessarily distinct) such that* rk $W = $ rk W', *and two points $\theta \in \mu^{-1}(W)$ and $\theta' \in \mu^{-1}(W')$ that both have the same rank list. Then there is a linear bijection from $\mathbb{R}^{d_\theta}$ to $\mathbb{R}^{d_\theta}$ that maps θ to θ', $\mu^{-1}(W)$ to $\mu^{-1}(W')$, and $S_{\underline{s}}^W$ to $S_{\underline{s}}^{W'}$ for every stratum $S_{\underline{s}}^W$ in the rank stratification of $\mu^{-1}(W)$.*

Proof. The function η is a linear bijection from $\mathbb{R}^{d_\theta}$ to $\mathbb{R}^{d_\theta}$ that maps $\tilde{\theta}$ to θ and, by Lemma 13, maps $\mu^{-1}(\tilde{I})$ to $\mu^{-1}(W)$ and $S_{\underline{s}}^{\tilde{I}}$ to $S_{\underline{s}}^W$. Likewise, there is a function η' that is a linear bijection from $\mathbb{R}^{d_\theta}$ to $\mathbb{R}^{d_\theta}$ that maps $\tilde{\theta}$ to θ', $\mu^{-1}(\tilde{I})$ to $\mu^{-1}(W')$, and $S_{\underline{s}}^{\tilde{I}}$ to $S_{\underline{s}}^{W'}$. Hence, $\eta' \circ \eta^{-1}$ is a linear bijection from $\mathbb{R}^{d_\theta}$ to $\mathbb{R}^{d_\theta}$ that maps θ to θ', $\mu^{-1}(W)$ to $\mu^{-1}(W')$, and $S_{\underline{s}}^W$ to $S_{\underline{s}}^{W'}$. $\square$

If we choose $W = W'$, Corollary 14 describes several automorphisms: a bijection that maps $\mu^{-1}(W)$ to itself and, for every stratum $S_{\underline{s}}^W$ in the rank stratification of $\mu^{-1}(W)$, a bijection that maps $S_{\underline{s}}^W$ to itself. The following corollary follows by letting $\underline{s}$ be the rank list $\underline{r}$ of θ.

Corollary 15. *For any two points θ, θ' in the same stratum $S_{\underline{r}}$, there is a linear homeomorphism from $S_{\underline{r}}$ to itself that maps θ to θ'.*

Corollary 15 implies that the stratum looks topologically the same at every point—that is, given an open neighborhood of one point on the stratum, every point on the stratum has an open neighborhood homeomorphic to that one. A consequence is that there are only two possibilities: every point on $S_{\underline{r}}$ has an open neighborhood homeomorphic to an open ball of some fixed dimension—hence $S_{\underline{r}}$ is a manifold—or no point on $S_{\underline{r}}$ has an open neighborhood homeomorphic to an open ball. The latter possibility is ruled out because every nonempty semi-algebraic set has at least one point with an open neighborhood homeomorphic to a ball.

We give some background on stratifications to justify this claim. Every semi-algebraic set can be stratified—that is, partitioned into strata that are manifolds without boundary—in a manner that satisfies three criteria. First, every stratum is analytic and thus smooth—specifically, every stratum is a manifold of class C^∞.

Second, the stratification is *locally finite*, meaning that every point of the semi-algebraic set has an open neighborhood that intersects only finitely many strata. Third, the stratification satisfies the frontier condition we defined in Section 3: for every pair of distinct strata S, T in the stratification, either $S \cap \bar{T} = \emptyset$ or $S \subseteq \bar{T}$. See Benedetti and Risler [3] for a proof. (Such stratifications were introduced for algebraic varieties by Whitney [22–24], though it was Mather [16] who showed that Whitney's stratifications satisfy the frontier condition, and Łojasiewicz [13] who generalized the result to semi-algebraic and semi-analytic sets. Thom [19] introduced the terms *stratum* and *stratification*. See Lu [15, Chapter 5] for an excellent exposition.) Here we use this result to stratify each stratum in the rank stratification—that is, $S_{\underline{r}}$ has a stratification satisfying the three criteria.

Let Z be a semi-algebraic set. Let $\mathcal{S}_Z$ be a locally finite stratification of Z satisfying the frontier condition and having C^∞-smooth strata. Let d be the maximum dimension among the strata in $\mathcal{S}_Z$. The *dimension of Z* is d. (This is by definition, though there are competing, equivalent definitions.) Let $S \in \mathcal{S}_Z$ be a stratum of dimension d, and let θ be any point on S. As S is a d-manifold, there is an open neighborhood $N \subset S$ of θ homeomorphic to an open d-ball. The frontier condition implies that no point in N lies in the closure of any stratum besides S (as S has the maximum dimension). The fact that $\mathcal{S}_Z$ is locally finite implies that no point in N lies in the closure of $Z \setminus S$. Therefore, N is an open set in Z (which is a stronger statement than it being an open set in S). Hence θ has an open neighborhood $N \subset Z$ homeomorphic to a d-ball.

Applying this knowledge with $Z = S_{\underline{r}}$, some point on $S_{\underline{r}}$ has an open neighborhood homeomorphic to a ball. Hence Corollary 15 implies that every point on $S_{\underline{r}}$ has an open neighborhood homeomorphic to a ball of the same dimension. We conclude that each stratum $S_{\underline{r}}$ in a rank stratification is a manifold.

Moreover, as $\theta \in N \subset S \subseteq S_{\underline{r}}$ and S is a manifold of class C^∞, the manifold $S_{\underline{r}}$ is C^∞-smooth at θ. Corollary 14 states that for every point $\theta' \in S_{\underline{r}}$, there is a linear homeomorphism from $S_{\underline{r}}$ to itself that maps θ to θ', so $S_{\underline{r}}$ is C^∞-smooth everywhere. Thus we have proven our main theorem.

Theorem 16. *Each stratum $S_{\underline{r}}$ of a fiber $\mu^{-1}(W)$ is a manifold of class C^∞.*

Unfortunately, this reasoning tells us nothing about the dimension of that manifold. Determining that dimension will require a good deal more work, which we undertake in Sections 8 and 9.

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

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

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 $\Delta W_{j\sim i} = W'_{j\sim i} - W_{j\sim i}$.

Moves on the fiber offer us a way to replace a linear neural network with another that computes the same function, but might be superior in other respects (such as not being near a spurious critical point of the cost function used to train the network). Moves on the fiber are also a tool for gaining intuition about the geometry and topology of the fiber. In Section 7, we use moves to understand how the strata in the rank stratification are connected to each other. In Section 8, we study the subspace tangent to a stratum (in weight space) and build a revealing basis for that tangent space; the moves described here will guide that study.

Two classes of move suffice to characterize strata and their interconnections: one-matrix moves and two-matrix moves. Any move from any point on the fiber to any other can be broken down into a sequence of one-matrix 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 $z \neq j$. 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 ($W' = W$, $\Delta W = 0$), 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.

  A one-matrix move stays on the fiber if $W_{L \sim j} \Delta W_j W_{j-1 \sim 0} = 0$; otherwise, it moves off the fiber. Some one-matrix moves change the ranks of one or more subsequence matrices; we call them *combinatorial moves*. A combinatorial move implies that θ' has a different rank list (and a different multiset of intervals) than θ. Some combinatorial moves stay on the fiber, and some move off of it, but all combinatorial moves move off of the stratum—that is, θ' does not lie on the stratum of the rank stratification that θ lies on (even if θ' lies on the same fiber). It is noteworthy that if two strata in the rank stratification satisfy $S_{\underline{r}} \cap \bar{S}_{\underline{s}} \neq \emptyset$, then from any point on $S_{\underline{r}}$, there is an infinitesimal combinatorial one-matrix move to $S_{\underline{s}}$; we will see that combinatorial one-matrix moves suffice to help us characterize all the connections among strata. We discuss one-matrix moves further in Section 6.1, and we study combinatorial moves in detail in Sections 7.3–7.7.

- A *two-matrix move* has exactly two nonzero displacement matrices ΔW_{j+1} and ΔW_j (which are always consecutive). That is, $W'_z = W_z$ for all $z \notin \{j, j+1\}$. We specify a two-matrix move by selecting an invertible $d_j \times d_j$ matrix M and setting $W'_{j+1} = W_{j+1} M$ and $W'_j = M^{-1} W_j$.

  Clearly, all two-matrix moves stay on the fiber: $W' = \mu(\theta') = \mu(\theta) = W$. Moreover, all two-matrix moves stay on the same stratum, as a two-matrix move does not change the rank of any subsequence matrix. (There are no combinatorial two-matrix moves.) 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.

  Unlike in a one-matrix move, often there is no straight path on the fiber from θ to θ'. But if M has a positive determinant, usually there is a natural choice of a smooth, curved path that lies on the fiber and on the stratum that contains θ. (By "smooth," we mean that every point on the path has a single well-defined tangent line.) Of particular interest to us is the direction tangent to that *two-matrix path* at θ, because that direction is also tangent to the stratum at θ.

  The fiber $\mu^{-1}(W)$ is not necessarily connected. In that case, it possible to move from one connected component of the fiber to another by means of a two-matrix move where M has a negative determinant (though there is no path on the fiber connecting θ to θ'). Section 6.2 discusses two-matrix moves.

6.1 One-Matrix Moves

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

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

where $W'_j = W_j + \Delta W_j$. Then

$$W' = \mu(\theta') = W_{L \sim j}(W_j + \Delta W_j)W_{j-1 \sim 0} = \mu(\theta) + W_{L \sim j}\Delta W_j W_{j-1 \sim 0} = W + W_{L \sim j}\Delta W_j W_{j-1 \sim 0}.$$

Therefore, θ' lies on the fiber $\mu^{-1}(W)$ if and only if ΔW_j satisfies $W_{L \sim j} \Delta W_j W_{j-1 \sim 0} = 0$. The set of displacements that satisfy this identity is the subspace (of $\mathbb{R}^{d_j \times d_{j-1}}$)

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

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

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

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 the dimension of N_j is

$$
\begin{aligned}
\dim N_j \;&=\; \dim\left(\text{null } W_{L \sim j} \otimes \mathbb{R}^{d_{j-1}}\right) + \dim\left(\mathbb{R}^{d_j} \otimes \text{null } W_{j-1 \sim 0}^{\top}\right) - \dim\left(\text{null } W_{L \sim j} \otimes \text{null } W_{j-1 \sim 0}^{\top}\right) \\
&=\; (d_j - \text{rk } W_{L \sim j}) \cdot d_{j-1} + d_j \cdot (d_{j-1} - \text{rk } W_{j-1 \sim 0}) - (d_j - \text{rk } W_{L \sim j}) \cdot (d_{j-1} - \text{rk } W_{j-1 \sim 0}) \\
&=\; d_j d_{j-1} - \text{rk } W_{L \sim j} \cdot \text{rk } W_{j-1 \sim 0}. \hspace{4cm} (6.1)
\end{aligned}
$$

Consider the affine subspace of the weight space $\mathbb{R}^{d_\theta}$ reachable from θ by a one-matrix move with $\Delta W_j \in N_j$,

$$
\zeta_j = \{(W_L, W_{L-1}, \ldots, W_{j+1}, W_j + \Delta W_j, W_{j-1}, \ldots, W_1) : \Delta W_j \in N_j\}.
$$

Then we have $\zeta_j \subseteq \mu^{-1}(W)$. That is, ζ_j is an affine subspace that is a subset of the fiber (but not necessarily a subset of the stratum that contains θ). This gives us some insight into the geometry of the fiber, but it tells us nothing about the curved parts of the fiber. The curved parts are revealed by the two-matrix moves.

6.2   Two-Matrix Moves and Two-Matrix Paths

Recall that a *two-matrix move* is a move that, for some $j \in [1, L-1]$, selects an invertible $d_j \times d_j$ matrix M and sets $W'_{j+1} = W_{j+1}M$ and $W'_j = M^{-1}W_j$ (but no other factor matrix changes). A two-matrix move always places θ' on the same fiber and the same stratum as θ. To see that, let $W = \mu(\theta)$ and let S be the stratum containing θ in the rank stratification of the fiber $\mu^{-1}(W)$. Observe that $W'_{k \sim i} = W_{k \sim i}$ when both $k \neq j$ and $i \neq j$. In particular, $W' = W$. The only subsequence matrices that change are those of the form $W'_{k \sim j} = W_{k \sim j}M$ where $k \in [j+1, L]$ and $W'_{j \sim i} = M^{-1}W_{j \sim i}$ where $i \in [0, j-1]$. But matrix rank is invariant to multiplication by an invertible matrix, so $\text{rk } W'_{k \sim i} = \text{rk } W_{k \sim i}$ for all k and i satisfying $L \geq k \geq i \geq 0$. Hence θ' has the same rank list as θ and thus lies on S.

Unlike a one-matrix move, a two-matrix move doesn't reveal a line on the fiber, but it can reveal a smooth, curved path on the fiber and on S, which tells us a direction tangent to S. Consider a two-matrix move where $M = I + \epsilon H$ for an arbitrary, nonzero $d_j \times d_j$ matrix H and a real $\epsilon > 0$. For a sufficiently small ϵ, M is invertible, so we can draw a smooth path on the fiber, with endpoint θ, by varying ϵ from zero to some small value. Thus we define the *two-matrix path*

$$
P = \{(W_L, W_{L-1}, \ldots, W_{j+2}, W_{j+1}(I + \epsilon H), (I + \epsilon H)^{-1}W_j, W_{j-1}, \ldots, W_1) : \epsilon \in [0, \hat{\epsilon}]\}. \hspace{1cm} (6.2)
$$

The curved grid lines in Figure 1 are examples of such paths. Every point on P is the target of some two-matrix move from θ, so $P \subseteq S$. To find the line tangent to P at θ, observe that for a small ϵ, $(I + \epsilon H)^{-1} = I - \epsilon H + \epsilon^2 H^2 - \epsilon^3 H^3 + \ldots$, so

$$\frac{\mathrm{d}}{\mathrm{d}\epsilon} W'_{j+1} = \frac{\mathrm{d}}{\mathrm{d}\epsilon}\left(W_{j+1}(I + \epsilon H)\right) = W_{j+1}H \quad \text{and}$$

$$\frac{\mathrm{d}}{\mathrm{d}\epsilon} W'_j\bigg|_{\epsilon=0} = \frac{\mathrm{d}}{\mathrm{d}\epsilon}\left((I + \epsilon H)^{-1} W_j\right)\bigg|_{\epsilon=0} = \frac{\mathrm{d}}{\mathrm{d}\epsilon}\left((I - \epsilon H + \epsilon^2 H^2 - \epsilon^3 H^3 + \ldots)W_j\right)\bigg|_{\epsilon=0} = -HW_j.$$

Let $T_\theta P$ be P's tangent line at θ and let $T_\theta S$ be S's tangent space at θ, both defined so that $T_\theta P$ and $T_\theta S$ pass through the origin, not necessarily through θ. Then

$$T_\theta P = \{(0, 0, \ldots, 0, \underbrace{\gamma W_{j+1}H}_{\Delta W_{j+1}}, \underbrace{-\gamma HW_j}_{\Delta W_j}, 0, \ldots, 0) : \gamma \in \mathbb{R}\}. \tag{6.3}$$

A key observation is that $T_\theta P \subseteq T_\theta S$, because $P \subseteq S$ and both P and S are smooth at θ.

In Section 8, we characterize $T_\theta S$ as a subspace spanned by some one-matrix and two-matrix tangent directions.

7 The One-Matrix Prebasis and the Hierarchy of Strata

In this section, we study how the strata are connected to each other. Strata form a partially ordered hierarchy in the following sense. Consider two nonempty strata $S_{\underline{r}}$ and $S_{\underline{s}}$ in a rank stratification of some fiber, with valid rank lists $\underline{r}$ and $\underline{s}$. Recall that $\underline{r} \leq \underline{s}$ means that $r_{k \sim i} \leq s_{k \sim i}$ for all $L \geq k \geq i \geq 0$, and $\underline{r} < \underline{s}$ means that $\underline{r} \leq \underline{s}$ and $\underline{r} \neq \underline{s}$ (at least one of the inequalities holds strictly). In Section 7.8, we will show that the following statements are equivalent (imply each other).

A. $S_{\underline{r}} \subseteq \bar{S}_{\underline{s}}$.
B. $S_{\underline{r}} \cap \bar{S}_{\underline{s}} \neq \emptyset$.
C. $\underline{r} \leq \underline{s}$.

The fact that A implies B is obvious (as we assume $S_{\underline{r}} \neq \emptyset$). More interesting is the fact that B implies A. That is, our stratifications satisfy the frontier condition defined in Section 3: if a stratum $S_{\underline{r}}$ intersects the closure of another stratum $S_{\underline{s}}$ from the same fiber, then $S_{\underline{r}}$ is a subset of $\bar{S}_{\underline{s}}$.

The most important observation is that $S_{\underline{r}} \subseteq \bar{S}_{\underline{s}}$ if and only if $\underline{r} \leq \underline{s}$. This makes it easy to determine which strata's closures include (or intersect) which strata. In Section 7.8 we will augment A, B, and C with a fourth equivalent claim, which permits us to arrange the strata in a directed acyclic graph (dag) like those illustrated in Figures 2, 3, and 4. Each vertex of the dag represents a stratum, and each directed edge of the dag represents a simple, tiny move called a *combinatorial move*; see Section 7.4. Every inclusion of one stratum in the closure of another is represented by one or more directed paths in the dag.

To achieve these goals, given a point $\theta \in \mu^{-1}(W)$, we will construct a *one-matrix prebasis at θ*, a prebasis that spans the weight space $\mathbb{R}^{d_\theta}$. Note that whereas in Section 4 we constructed prebases for vectors, now we will construct prebases for matrices (tensor products of the vector prebases from Section 4) and a prebasis for weight vectors (sequences of matrices). These prebases are different at each weight vector $\theta \in \mathbb{R}^{d_\theta}$.

Each member of the one-matrix prebasis is a *one-matrix subspace* whose vectors are one-matrix move directions, defined in Section 6.1. Some of these subspaces represent moves that stay on the fiber, and some

represent moves that move off the fiber. Some of the subspaces represent moves that move to different strata. All the strata that meet at θ can be accessed from θ by one or more infinitesimal one-matrix moves whose displacements are in these one-matrix subspaces. Each move corresponds to an edge in the dag.

Another purpose of the one-matrix prebasis is to help identify the space tangent to a stratum S at a point $\theta \in S$, denoted $T_\theta S$. Specifically, we construct a prebasis for $T_\theta S$. We are not satisfied with that alone; for every stratum S' whose closure contains θ, we would like to identify the tangent space $T_\theta S'$ (which is a superset of $T_\theta S$) if it exists. Moreover, we want a *single* prebasis that can express *all* of these tangent spaces at θ. (The prebasis that does this is called the *fiber prebasis*, described in Section 8.) This information allows us to identify all the directions from θ of paths that leave S and enter various higher-dimensional strata.

To visualize this, recall Figure 3 and consider a point θ on the stratum S_{001}. The space tangent to S_{001} at θ is a line; the one-matrix prebasis at θ contains this line. S_{001} lies in the closures of two two-dimensional strata, S_{011} and S_{101}. The space tangent to S_{011} at θ is a plane, which we can represent as the vector sum of the same line we used for S_{001} and one additional line (not uniquely defined). The space tangent to S_{101} at θ is also a plane, which we can represent with the same line we used for S_{001} plus a different additional line. The set of these three lines is a one-matrix prebasis at θ. The lines reflect three degrees of freedom by which paths on the fiber can leave θ (though no single path can exploit more than two degrees of freedom).

Some one-matrix moves change the ranks of one or more subsequence matrices; we call them *combinatorial moves*. If a combinatorial move stays on the fiber, it moves to a different stratum of the fiber. These moves are our main source of insight into how strata are connected to each other; we study them in Sections 7.3–7.7. The one-matrix prebasis distinguishes moves that increase the rank of one or more subsequence matrices from moves that do not, and it also distinguishes moves according to *which* subsequence matrices have their ranks increased. (It does not distinguish moves that *decrease* the rank of one or more subsequence matrices from moves that do not.) Note that there are no combinatorial two-matrix moves; two-matrix moves do not change the rank of any subsequence matrix.

The one-matrix prebasis suffices to characterize all the interconnections among strata, as we will see in Section 7.8. But the one-matrix prebasis does not give us all the subspaces we need to characterize the tangent space $T_\theta S$. To account for directions along which S is curved, we will define some *two-matrix subspaces* in Section 8. A prebasis spanning $T_\theta S$ will enable us to determine the dimension of S.

7.1 Small Moves

Recall that a move is *combinatorial* if it changes the rank of one or more of the subsequence matrices (thereby moving to a different stratum). Here we characterize what we call *small moves*, which are motivated by the fact that an infinitesimal perturbation of a matrix can increase its rank, but cannot decrease its rank—given a matrix, there is a positive lower bound on the magnitude of a displacement that can decrease its rank. By studying moves that are so small that no subsequence matrix can decrease in rank, we simplify understanding the interconnections among strata. 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 them lie in the closure of the 2-dimensional stratum S_{011}. Starting from any point $\theta \in S_{010}$, an infinitesimally small move can reach some point $\theta' \in S_{011}$ (because θ lies in the closure of S_{011}), which entails an increase in the rank of W_1 from 0 to 1. But from θ, an infinitesimally small move does not suffice to reach S_{000}, as θ is not in the closure of S_{000}.

If a stratum $S_{\underline{r}}$ intersects the closure of a stratum $S_{\underline{s}}$, then an infinitesimal move can get from $S_{\underline{r}}$ to $S_{\underline{s}}$. But an infinitesimally small perturbation cannot decrease a matrix's rank. Decreasing the rank of a subsequence matrix always entails moving some finite distance. This implies that $\underline{r} \leq \underline{s}$. We prove it formally.

Lemma 17. *If a stratum $S_{\underline{r}}$ intersects the closure of a stratum $S_{\underline{s}}$ from the same fiber, then $\underline{r} \leq \underline{s}$.*

Proof. By assumption, there exists a point $\theta \in S_{\underline{r}} \cap \bar{S}_{\underline{s}}$. Therefore, every open neighborhood $N \subset \mathbb{R}^{d_\theta}$ of θ intersects $S_{\underline{s}}$. But there exists an open neighborhood $N \subset \mathbb{R}^{d_\theta}$ of θ such that for every weight vector $\theta_N \in N$, the rank list $\underline{t}$ of θ_N satisfies $r_{k\sim i} \leq t_{k\sim i}$ for all $L \geq k \geq i \geq 0$. As N intersects $S_{\underline{s}}$, there exists a point $\theta' \in N \cap S_{\underline{s}}$. The rank list of θ' is $\underline{s}$, so $r_{k\sim i} \leq s_{k\sim i}$ for all $L \geq k \geq i \geq 0$. $\qquad \square$

In terms of the equivalent statements we made at the start of Section 7, Lemma 17 states that B implies C. In Section 7.8, we will prove that C implies A, thereby establishing that $S_{\underline{r}} \subseteq \bar{S}_{\underline{s}}$ if and only if $\underline{r} \leq \underline{s}$.

Infinitesimals have a complicated status in the history of mathematical rigor. To strip away everything that is not essential, we define a *small move* to be any move such that every stratum-closure that contains the destination point θ' also contains the starting point θ. That is, you can never enter a stratum's closure by a small move if you're not already there. This definition has a counterintuitive consequence: the inverse of a small move is *not* necessarily small. Specifically, if a small move increases the rank of some subsequence matrix, its inverse is not small. Moving from S_{010} to S_{011} is small, but moving back is not. It follows from the definition of "small" that if a small move moves from a stratum $S_{\underline{r}}$ to a stratum $S_{\underline{s}}$, then $S_{\underline{r}} \cap \bar{S}_{\underline{s}} \neq \emptyset$, so $\underline{r} \leq \underline{s}$ by Lemma 17.

The *small combinatorial moves* are the small moves that increase some subsequence matrix rank (as they cannot decrease any matrix rank). A small combinatorial move from $S_{\underline{r}}$ to $S_{\underline{s}}$ implies that $\underline{r} < \underline{s}$ and thus $S_{\underline{r}}$ is disjoint from $S_{\underline{s}}$. But we will see that $S_{\underline{r}} \subseteq \bar{S}_{\underline{s}}$, so a small combinatorial move always moves to a stratum of higher dimension. The small combinatorial moves establish a natural partial ordering of the strata that reflects the inclusions among stratum-closures.

7.2 The One-Matrix Subspaces and the One-Matrix Prebasis

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—and in Section 4.6 we decompose it further into a basis of vectors (a more familiar concept). Here we construct prebases for the factor matrix spaces $\mathbb{R}^{d_j \times d_{j-1}}$ and the weight space $\mathbb{R}^{d_\theta}$. Given a starting point $\theta \in \mathbb{R}^{d_\theta}$, one of our goals is that our prebases should separate one-matrix moves that stay on the fiber from those that do not. Moreover, they should separate one-matrix moves that stay on the closure of the stratum from those that do not, and they should do so for every stratum whose closure contains θ. We saw a hint about how to achieve that in Section 6.1, where we expressed the set of displacements ΔW_j that stay on the fiber as $N_j = A_{L-1,j,j} \otimes \mathbb{R}^{d_{j-1}} + \mathbb{R}^{d_j} \otimes B_{j-1,j-1,1}$, in terms of the flow subspaces defined in Section 4.2.

First, we construct a prebasis O_j for $\mathbb{R}^{d_j \times d_{j-1}}$. For indices 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

$$o_{lkjih} = a_{lji} \otimes b_{k,j-1,h} = \{M \subseteq \mathbb{R}^{d_j \times d_{j-1}} : \operatorname{col} M \subseteq a_{lji} \text{ and } \operatorname{row} M \subseteq b_{k,j-1,h}\},$$

where a_{lji} and $b_{k,j-1,h}$ are prebasis subspaces, as defined in Section 4.3. Recall that these subspaces depend on the starting point θ. Lemma 7 guarantees that we can choose flow subspaces (i.e., subspaces such that $a_{kji} = W_{j\sim x} a_{kxi}$ and $b_{kji} = W_{y\sim j}^\top b_{kyi}$); for some of our results here, it is necessary that we do so.

Recall from Sections 4.6 and 4.7 the matrix J_{lji} whose ω_{li} columns are a basis for a_{lji} and the matrix $K_{k,j-1,h}$ whose ω_{kh} columns are a basis for $b_{k,j-1,h}$. We can easily construct a basis for o_{lkjih}, namely, the set

$$\{uv^\top : u \text{ is a column of } J_{lji} \text{ and } v \text{ is a column of } K_{k,j-1,h}\}.$$

This basis is composed of $\omega_{li} \omega_{kh}$ rank-1 matrices. Hence another definition of the subspace o_{lkjih} is

$$o_{lkjih} = \{J_{lji} C K_{k,j-1,h}^\top : C \in \mathbb{R}^{\omega_{li} \times \omega_{kh}}\},$$

where the matrix C holds the coefficients of the basis matrices. It is clear 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

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

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

Lemma 18. O_j *is a prebasis for* $\mathbb{R}^{d_j \times d_{j-1}}$. *In particular, the subspaces in* O_j *are linearly independent.*

Proof. The vector sum of the one-matrix subspaces in O_j is

$$\sum_{o_{lkjih} \in O_j} o_{lkjih} \;=\; \sum_{a_{lji} \in \mathcal{A}_j} \sum_{b_{k,j-1,h} \in \mathcal{B}_{j-1}} a_{lji} \otimes b_{k,j-1,h} = \left(\sum_{a_{lji} \in \mathcal{A}_j} a_{lji}\right) \otimes \left(\sum_{b_{k,j-1,h} \in \mathcal{B}_{j-1}} b_{k,j-1,h}\right)$$

$$= \;\; \mathbb{R}^{d_j} \otimes \mathbb{R}^{d_{j-1}} = \mathbb{R}^{d_j \times d_{j-1}}.$$

(The second line follows from Lemma 4.) Hence, O_j spans $\mathbb{R}^{d_j \times d_{j-1}}$, which has dimension $d_j d_{j-1}$. The sum of the dimensions of these subspaces is

$$\sum_{o_{lkjih} \in O_j} \omega_{li}\,\omega_{kh} = \sum_{a_{lji} \in \mathcal{A}_j} \sum_{b_{k,j-1,h} \in \mathcal{B}_{j-1}} \omega_{li}\,\omega_{kh} = \left(\sum_{a_{lji} \in \mathcal{A}_j} \omega_{li}\right) \cdot \left(\sum_{b_{k,j-1,h} \in \mathcal{B}_{j-1}} \omega_{kh}\right) = d_j d_{j-1}.$$

As $\mathbb{R}^{d_j \times d_{j-1}}$ is the vector sum of the subspaces and its dimension equals the sum of the dimensions of the subspaces, the subspaces are linearly independent. Hence O_j is a prebasis. □

Now we construct a prebasis Θ_O for $\mathbb{R}^{d_\theta}$ that we call the *one-matrix prebasis*, the main topic of this section. The subspaces in Θ_O are called the *one-matrix subspaces* and have the form

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

with M in position j from the right (the same position as W_j in θ). The one-matrix prebasis is

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

It is easy to see that Θ_O is a prebasis for $\mathbb{R}^{d_\theta}$ as a corollary of Lemma 18.

Corollary 19. Θ_O *is a prebasis for* $\mathbb{R}^{d_\theta}$. *In particular, the subspaces in* Θ_O *are linearly independent.*

Proof. For any fixed $j \in [1, L]$, Θ_O contains a subspace $\phi_{lkjih} = \{(0, \ldots, 0, M, 0, \ldots 0) : M \in o_{lkjih}\}$ for every choice of the four indices $l \in [j, L]$, $k \in [j - 1, L]$, $i \in [0, j]$, and $h \in [0, j - 1]$, where M occurs at position j in the weight vector. The set O_j contains one subspace o_{lkjih} for every $l \in [j, L]$, $k \in [j - 1, L]$, $i \in [0, j]$, and $h \in [0, j - 1]$. By Lemma 18, O_j is a prebasis for $\mathbb{R}^{d_j \times d_{j-1}}$. Hence, for a fixed j and varying l, k, i, and h, $\sum_{\phi_{lkjih} \in \Theta_O} \phi_{lkjih} = \{(0, \ldots, 0, M, 0, \ldots 0) : M \in \mathbb{R}^{d_j \times d_{j-1}}\}$. If we sum over $j \in [1, L]$ as well, we have $\sum_{\phi_{lkjih} \in \Theta_O} \phi_{lkjih} = \mathbb{R}^{d_\theta}$. The sum of the dimensions of the subspaces in Θ_O is $\sum_{j=1}^{L} d_j d_{j-1} = d_\theta$, matching the dimension of $\mathbb{R}^{d_\theta}$, so the subspaces in Θ_O are linearly independent. Therefore, Θ_O is a prebasis for $\mathbb{R}^{d_\theta}$. □

We can distinguish the one-matrix subspaces into two types: those that when translated to pass through θ are subsets of the fiber, and those that when translated to pass through θ intersect the fiber at only one point (namely, θ). Hence, moves with displacements in the former subspaces stay on the fiber, and moves with nonzero displacements in the latter subspaces move off the fiber. This motivates us to partition O_j into two sets, O_j^{fiber} and O_j^{L0}, and to partition Θ_O into two sets, Θ_O^{fiber} and Θ_O^{L0}.

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

For example, in the two-matrix case ($L = 2$),

$$
\begin{aligned}
O_1^{20} &= \{o_{20100}, o_{20110}, o_{21100}, o_{21110}, o_{22100}, o_{22110}\} \setminus \{\{\mathbf{0}\}\}, \\
O_2^{20} &= \{o_{21200}, o_{21210}, o_{21220}, o_{22200}, o_{22210}, o_{22220}\} \setminus \{\{\mathbf{0}\}\}, \\
\Theta_O^{20} &= \{\phi_{20100}, \phi_{20110}, \phi_{21100}, \phi_{21110}, \phi_{22100}, \phi_{22110}, \phi_{21200}, \phi_{21210}, \phi_{21220}, \phi_{22200}, \phi_{22210}, \phi_{22220}\} \setminus \{\{\mathbf{0}\}\}, \\
O_1^{\text{fiber}} &= \{o_{10100}, o_{10110}, o_{11100}, o_{11110}, o_{12100}, o_{12110}\} \setminus \{\{\mathbf{0}\}\}, \\
O_2^{\text{fiber}} &= \{o_{21201}, o_{21211}, o_{21221}, o_{22201}, o_{22211}, o_{22221}\} \setminus \{\{\mathbf{0}\}\}, \\
\Theta_O^{\text{fiber}} &= \{\phi_{10100}, \phi_{10110}, \phi_{11100}, \phi_{11110}, \phi_{12100}, \phi_{12110}, \phi_{21201}, \phi_{21211}, \phi_{21221}, \phi_{22201}, \phi_{22211}, \phi_{22221}\} \setminus \{\{\mathbf{0}\}\}.
\end{aligned}
$$

We end this section by proving three things.

- O_j^{fiber} is a prebasis for N_j, where N_j is defined in Section 6.1. (Lemma 20 below.)

- Θ_O^{fiber} contains the one-matrix subspaces that "stay on the fiber." Specifically, every displacement $\Delta\theta \in \phi_{lkjih}$ with $\phi_{lkjih} \in \Theta_O^{\text{fiber}}$ satisfies $\mu(\theta + \Delta\theta) = W$. (Corollary 21 below.)

- Θ_O^{L0} contains the one-matrix subspaces that "move off the fiber." Specifically, every displacement $\Delta\theta \in \phi_{lkjih} \setminus \{\mathbf{0}\}$ with $\phi_{lkjih} \in \Theta_O^{L0}$ satisfies $\mu(\theta + \Delta\theta) \neq W$. (Also Corollary 21.)

Henceforth, for any set of subspaces Θ, we define the *span* of Θ to be $\text{span}\,\Theta = \sum_{\phi \in \Theta} \phi$, the vector sum of the subspaces in Θ.

Lemma 20. 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,

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

Therefore,

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

By Lemma 18, 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. □

Corollary 21. *For every subspace $\phi_{lkjih} \in \Theta_O^{\text{fiber}}$ and every displacement $\Delta\theta \in \phi_{lkjih}$, $\mu(\theta + \Delta\theta) = W$. For every subspace $\phi_{lkjih} \in \Theta_O^{L0}$ and every displacement $\Delta\theta \in \phi_{lkjih} \setminus \{0\}$, $\mu(\theta + \Delta\theta) \neq W$.*

Proof. Every subspace $\phi_{lkjih} \in \Theta_O$ is a one-matrix subspace, so a displacement $\Delta\theta \in \phi_{lkjih}$ has at most one nonzero matrix, $\Delta W_j \in o_{lkjih}$. Recall that $\mu(\theta + \Delta\theta) = W + W_{L\sim j}\Delta W_j W_{j-1\sim 0}$. If $\phi_{lkjih} \in \Theta_O^{\text{fiber}}$, then $o_{lkjih} \in O_j^{\text{fiber}}$ and thus $o_{lkjih} \subseteq N_j$ by Lemma 20, so $\Delta W_j \in N_j$ and $\mu(\theta + \Delta\theta) = W$.

By constrast, if $\phi_{lkjih} \in \Theta_O^{L0}$ and $\Delta\theta \in \phi_{lkjih} \setminus \{0\}$, then $o_{lkjih} \in O_j^{L0} = O_j \setminus O_j^{\text{fiber}}$ and $\Delta W_j \in o_{lkjih} \setminus \{0\}$. By Lemma 18, the subspaces in O_j are linearly independent, so $o_{lkjih} \cap \operatorname{span} O_j^{\text{fiber}} = \{0\}$. Then by Lemma 20, $o_{lkjih} \cap N_j = \{0\}$ and thus $\Delta W_j \notin N_j$ and $\mu(\theta + \Delta\theta) \neq W$. □

7.3 The Effects of One-Matrix Moves with Displacements in the One-Matrix Prebasis

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' has a different rank list than θ (by definition), θ' has a different multiset of intervals than θ, and crucially, θ' is in a different stratum than θ, of a different dimension.

For each subspace o_{lkjih} in the prebasis 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 change rank? Which subsequence matrices undergo a change in rowspace or columnspace? This section answers these questions. Table 5 summarizes the answers for small moves. The answers will clarify why we chose subspaces of the form $o_{lkjih} = a_{lji} \otimes b_{k,j-1,h}$.

To set up every lemma in this section, let $\Delta W_j = \epsilon M$ for a scalar $\epsilon \in \mathbb{R}$ and a matrix $M \in o_{lkjih} \setminus \{0\}$. Assume $l \geq j \geq i$ and $k \geq j-1 \geq h$ (otherwise o_{lkjih} 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 22. *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 $x < h$. Moreover, if none of those four conditions holds, then $\operatorname{rk} \Delta W_{y\sim x} = \operatorname{rk} \Delta W_j$.*

(a) swapping move

Axes: y (vertical, with W), x (horizontal). Vertical labels top-to-bottom: L, $l+1$, l, $k+1$, k, j, $j-1$, 0. Horizontal labels: 0, $h-1$, h, $i-1$, i, $j-1$, j, L.

	0–$h-1$	h–$i-1$	i–$j-1$	j–L
$l+1$–L	$W'_{y\sim x} = W_{y\sim x}$ (blue)	$W'_{y\sim x} = W_{y\sim x}$ (blue)	$W'_{y\sim x} = W_{y\sim x}$ (blue)	$W'_{y\sim x} = W_{y\sim x}$ (red)
$k+1$–l	$W'_{y\sim x} = W_{y\sim x}$ (blue)	$W'_{y\sim x} \neq W_{y\sim x}$, $\operatorname{rk} W'_{y\sim x} = \operatorname{rk} W_{y\sim x} + \operatorname{rk}\Delta W_j$, $\operatorname{col} W'_{y\sim x} \nsubseteq \operatorname{col} W_{y\sim x}$, $\operatorname{row} W'_{y\sim x} \nsubseteq \operatorname{row} W_{y\sim x}$ (yellow)	$W'_{y\sim x} \neq W_{y\sim x}$, $\operatorname{rk} W'_{y\sim x} = \operatorname{rk} W_{y\sim x}$, $\operatorname{col} W'_{y\sim x} = \operatorname{col} W_{y\sim x}$, $\operatorname{row} W'_{y\sim x} \nsubseteq \operatorname{row} W_{y\sim x}$	$W'_{y\sim x} = W_{y\sim x}$ (red)
j–k	$W'_{y\sim x} = W_{y\sim x}$ (blue)	$W'_{y\sim x} \neq W_{y\sim x}$, $\operatorname{rk} W'_{y\sim x} = \operatorname{rk} W_{y\sim x}$, $\operatorname{col} W'_{y\sim x} \nsubseteq \operatorname{col} W_{y\sim x}$, $\operatorname{row} W'_{y\sim x} = \operatorname{row} W_{y\sim x}$	$W'_{y\sim x} \neq W_{y\sim x}$, $\operatorname{rk} W'_{y\sim x} = \operatorname{rk} W_{y\sim x}$, $\operatorname{col} W'_{y\sim x} = \operatorname{col} W_{y\sim x}$, $\operatorname{row} W'_{y\sim x} = \operatorname{row} W_{y\sim x}$	$W'_{y\sim x} = W_{y\sim x}$ (red)
0–$j-1$	$W'_{y\sim x} = W_{y\sim x}$ (red)	$W'_{y\sim x} = W_{y\sim x}$ (red)	$W'_{y\sim x} = W_{y\sim x}$ (red)	

(W_j arrow points to hatched region.)

(b) connecting move

Axes: y (with W), x. Vertical labels: L, $l+1$, l, j, $k=j-1$, 0. Horizontal labels: 0, $h-1$, h, $j-1$, $i=j$, L.

- $l+1$–L: $W'_{y\sim x} = W_{y\sim x}$ (blue)
- l–l (middle yellow box): $W'_{y\sim x} \neq W_{y\sim x}$, $\operatorname{rk} W'_{y\sim x} = \operatorname{rk} W_{y\sim x} + \operatorname{rk}\Delta W_j$, $\operatorname{col} W'_{y\sim x} \nsubseteq \operatorname{col} W_{y\sim x}$, $\operatorname{row} W'_{y\sim x} \nsubseteq \operatorname{row} W_{y\sim x}$ (yellow)
- right column: $W'_{y\sim x} = W_{y\sim x}$ (red)
- $k=j-1$ row: $W'_{y\sim x} = W_{y\sim x}$ (red)

(W_j arrow points to hatched region.)

(c) not a combinatorial move

Axes: y (with W), x. Vertical labels: L, k, l, j, $j-1$, 0. Horizontal labels: $h=i=0$, $j-1$, j, L.

- k–L: $W'_{y\sim x} = W_{y\sim x}$ (blue)
- l row: $W'_{y\sim x} \neq W_{y\sim x}$, $\operatorname{rk} W'_{y\sim x} = \operatorname{rk} W_{y\sim x}$, $\operatorname{col} W'_{y\sim x} = \operatorname{col} W_{y\sim x}$, $\operatorname{row} W'_{y\sim x} = \operatorname{row} W_{y\sim x}$
- right column: $W'_{y\sim x} = W_{y\sim x}$ (red)
- $j-1$ row: $W'_{y\sim x} = W_{y\sim x}$ (red)

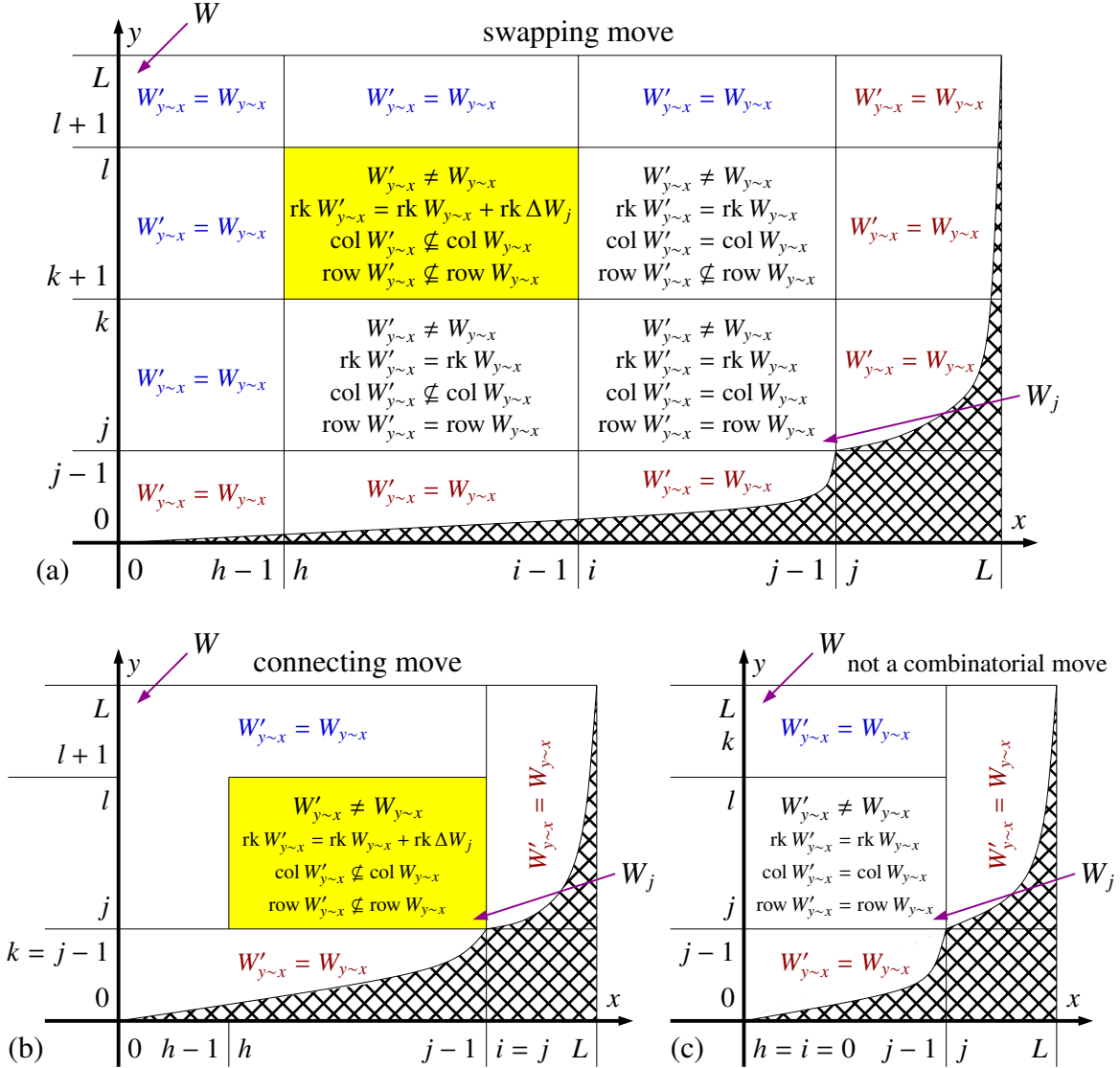(W_j arrow points to hatched region.)

Table 5: The influence of a one-matrix move in which a factor matrix W_j undergoes a sufficiently small, nonzero 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 $l = L$, though we don't depict that case here. If $h = 0$ and $l = L$, then $W' \neq W$ and we move off the fiber.)

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

Now consider the case where none of the four conditions holds—that is, the case where $\epsilon \neq 0$ and $j \in [x+1, y]$ and $y \leq l$ and $x \geq h$. By Lemma 6, $W_{y \sim j} a_{lji}$ has the same dimension as a_{lji}, and $W^{\top}_{j-1 \sim x} b_{k,j-1,h}$ has the same dimension as $b_{k,j-1,h}$. In other words, the application of $W_{y \sim j}$ is a bijection from a_{lji} to $W_{y \sim j} a_{lji}$, and the application of $W^{\top}_{j-1 \sim x}$ is a bijection from $b_{k,j-1,h}$ to $W^{\top}_{j-1 \sim x} b_{k,j-1,h}$. Hence rk $(W_{y \sim j} M) = \dim \operatorname{col}(W_{y \sim j} M) = \dim \operatorname{col} M = \operatorname{rk} M$ and rk $(M W_{j-1 \sim x}) = \dim \operatorname{row}(M W_{j-1 \sim x}) = \dim \operatorname{row} M = \operatorname{rk} M$. By the Frobenius rank inequality, rk $\Delta W_{y \sim x} = \operatorname{rk}(W_{y \sim j} M W_{j-1 \sim x}) \geq \operatorname{rk}(W_{y \sim j} M) + \operatorname{rk}(M W_{j-1 \sim x}) - \operatorname{rk} M = \operatorname{rk} M$. The rank of $W_{y \sim j} M W_{j-1 \sim x}$ cannot exceed rk M, so rk $\Delta W_{y \sim x} = \operatorname{rk} M = \operatorname{rk} \Delta W_j$ as claimed. By assumption, $M \neq 0$, so $\Delta W_{y \sim x} \neq 0$ and $W'_{y \sim x} \neq W_{y \sim x}$ as claimed. □

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

Lemma 23. *Given $L \geq y \geq x \geq 0$, col $\Delta W_{y \sim x} \subseteq$ col $W_{y \sim x}$ (equivalently, col $W'_{y \sim x} \subseteq$ col $W_{y \sim x}$) if and only if $\epsilon = 0$ or $j \notin [x+1, y]$ or $y > l$ or $x < h$ or $x \geq i$. Moreover, if none of those five conditions holds, then col $\Delta W_{y \sim x} \cap$ col $W_{y \sim x} = \{0\}$.*

Symmetrically, row $\Delta W_{y \sim x} \subseteq$ row $W_{y \sim x}$ (equivalently, row $W'_{y \sim x} \subseteq$ row $W_{y \sim x}$) if and only if $\epsilon = 0$ or $j \notin [x+1, y]$ or $y > l$ or $x < h$ or $y \leq k$. Moreover, if none of those five conditions holds, then row $\Delta W_{y \sim x} \cap$ row $W_{y \sim x} = \{0\}$.

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 22 and the results follow.

Henceforth, assume that $\epsilon \neq 0$ and $j \in [x+1, y]$ and $y \leq l$ and $x \geq h$. By Lemma 22, rk $\Delta W_{y \sim x} = \operatorname{rk} \Delta W_j = \operatorname{rk} M$. By assumption, $M \neq 0$; therefore $\Delta W_{y \sim x} \neq 0$, col $\Delta W_{y \sim x} \neq \{0\}$, and row $\Delta W_{y \sim x} \neq \{0\}$.

Recall that $\Delta W_{y \sim x} = \epsilon W_{y \sim j} M W_{j-1 \sim x}$ and col $M \subseteq a_{lji}$. Thus col $\Delta W_{y \sim x} \subseteq \operatorname{col}(W_{y \sim j} M) = W_{y \sim j} \operatorname{col} M \subseteq W_{y \sim j} a_{lji} \subseteq W_{y \sim j} A_{lji} \subseteq W_{y \sim j} \operatorname{col} W_{j \sim i} = \operatorname{col} W_{y \sim i}$. If $x \geq i$ then col $W_{y \sim i} \subseteq$ col $W_{y \sim x}$, so col $\Delta W_{y \sim x} \subseteq$ col $W_{y \sim x}$ as claimed.

Symmetrically, as row $M \subseteq b_{k,j-1,h}$, row $\Delta W_{y \sim x} \subseteq \operatorname{row}(M W_{j-1 \sim x}) = W^{\top}_{j-1 \sim x} \operatorname{row} M \subseteq W^{\top}_{j-1 \sim x} b_{k,j-1,h} \subseteq W^{\top}_{j-1 \sim x} B_{k,j-1,h} \subseteq W^{\top}_{j-1 \sim x} \operatorname{row} W_{k \sim j-1} = \operatorname{row} W_{k \sim x}$. If $y \leq k$ then row $W_{k \sim x} \subseteq$ row $W_{y \sim x}$, so row $\Delta W_{y \sim x} \subseteq$ row $W_{y \sim x}$ as claimed.

It remains to consider the cases where $x < i$ or $y > k$—consider the former first. Recall that col $W_{j \sim x} = A_{Ljx} = \sum_{z=j}^{L} \sum_{w=0}^{x} a_{zjw}$. If $x < i$, this summation does not include the term a_{lji}. By Lemma 4, $\mathcal{A}_j$ is a prebasis for $\mathbb{R}^{d_j}$ whose subspaces include a_{lji} and all the terms a_{zjw} in the summation. The subspaces in a prebasis are linearly independent; hence $a_{lji} \cap \operatorname{col} W_{j \sim x} = \{0\}$. Premultiplying both sides by $W_{y \sim j}$ gives $W_{y \sim j} a_{lji} \cap \operatorname{col} W_{y \sim x} = \{0\}$. Two paragraphs ago, we saw that col $\Delta W_{y \sim x} \subseteq W_{y \sim j} a_{lji}$. We conclude that if $x < i$, col $\Delta W_{y \sim x} \cap$ col $W_{y \sim x} = \{0\}$ and col $\Delta W_{y \sim x} \not\subseteq$ col $W_{y \sim x}$ as claimed.

Symmetrically, row $W_{y \sim j-1} = B_{y,j-1,0} = \sum_{z=y}^{L} \sum_{w=0}^{j-1} b_{z,j-1,w}$. If $y > k$, this summation does not include the term $b_{k,j-1,h}$. By Lemma 4, $\mathcal{B}_{j-1}$ is a prebasis for $\mathbb{R}^{d_{j-1}}$ whose subspaces include $b_{k,j-1,h}$ and all the terms $b_{z,j-1,w}$ in the summation. Hence $b_{k,j-1,h} \cap \operatorname{row} W_{y \sim j-1} = \{0\}$. Premultiplying both sides by $W^{\top}_{j-1 \sim x}$ gives $W^{\top}_{j-1 \sim x} b_{k,j-1,h} \cap \operatorname{row} W_{y \sim x} = \{0\}$. Two paragraphs ago, we saw that row $\Delta W_{y \sim x} \subseteq W^{\top}_{j-1 \sim x} b_{k,j-1,h}$. We conclude that if $y > k$, row $\Delta W_{y \sim x} \cap$ row $W_{y \sim x} = \{0\}$ and row $\Delta W_{y \sim x} \not\subseteq$ 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, then gives a stronger statement for moves that are sufficiently small.

Lemma 24. *Given $L \geq y \geq x \geq 0$, for all $\epsilon \in \mathbb{R}$, rk $W'_{y \sim x} \leq$ rk $W_{y \sim x} +$ rk ΔW_j. Moreover, if $y > l$ or $y \leq k$ or $x \geq i$ or $x < h$, then rk $W'_{y \sim x} \leq$ rk $W_{y \sim x}$.*

Moreover, there exists an $\hat{\epsilon} > 0$ such that for all $\epsilon \in (-\hat{\epsilon}, \hat{\epsilon})$,

$$\text{rk } W'_{y \sim x} = \begin{cases} \text{rk } W_{y \sim x} + \text{rk } \Delta W_j & \epsilon \neq 0 \text{ and } l \geq y > k \text{ and } i > x \geq h; \\ \text{rk } W_{y \sim x} & \text{otherwise.} \end{cases}$$

Proof. The displacement $\Delta W_{y \sim x} = W_{y \sim j} \Delta W_j W_{j-1 \sim x}$ has rank at most rk ΔW_j, so rk $W'_{y \sim x} \leq$ rk $W_{y \sim x} +$ rk $\Delta W_{y \sim x} \leq$ rk $W_{y \sim x} +$ rk ΔW_j. If $y > l$ or $y \leq k$, then row $W'_{y \sim x} \subseteq$ row $W_{y \sim x}$ by Lemma 23, so rk $W'_{y \sim x} \leq$ rk $W_{y \sim x}$. If $x \geq i$ or $x < h$, then col $W'_{y \sim x} \subseteq$ col $W_{y \sim x}$ by Lemma 23, and again rk $W'_{y \sim x} \leq$ 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 rk $W'_{y \sim x} =$ 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 23, col $\Delta W_{y \sim x} \cap$ col $W_{y \sim x} = \{\mathbf{0}\}$ and row $\Delta W_{y \sim x} \cap$ row $W_{y \sim x} = \{\mathbf{0}\}$. Therefore, if ϵ is sufficiently small, then rk $W'_{y \sim x} =$ rk $W_{y \sim x} +$ rk $\Delta W_{y \sim x}$. By Lemma 22, rk $\Delta W_{y \sim x} =$ rk ΔW_j, so rk $W'_{y \sim x} =$ rk $W_{y \sim x} +$ rk ΔW_j. $\qquad \square$

The second half of Lemma 24 applies to small moves. Table 5 illustrates the parts of Lemmas 22, 23, and 24 that apply to small moves with nonzero displacements.

7.4 Connecting Moves and Swapping Moves

Consider small moves from the one-matrix prebasis Θ_O—that is, moves that replace θ with $\theta' = \theta + \Delta\theta$, where $\Delta\theta \in \phi_{lkjih}$ is sufficiently small and $\phi_{lkjih} \in \Theta_O$. (Equivalently, moves that replace W_j with $W'_j = W_j + \Delta W_j$, where $\Delta W_j \in o_{lkjih}$ is sufficiently small and $o_{lkjih} \in O_j$). Lemma 24 shows that if $\Delta\theta \neq \mathbf{0}$, 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 5 illustrates). Their ranks all increase by the same amount: the rank of ΔW_j. Hence, a small move with displacement $\Delta\theta \in \phi_{lkjih}$ is combinatorial if and only if $l > k$, $i > h$, and $\Delta\theta \neq \mathbf{0}$.

Interestingly, although a single move may change the ranks of many subsequence matrices, at most four interval multiplicities change. Recall the identity (4.10), $\omega_{yx} =$ rk $W_{y \sim x} -$ rk $W_{y \sim x-1} -$ rk $W_{y+1 \sim x} +$ rk $W_{y+1 \sim x-1}$. If all four ranks increase by rk ΔW_j, or exactly two ranks with opposite signs do, then ω_{yx} does not change.

It is straightforward to check that ω_{kh} and ω_{li} decrease by rk ΔW_j, ω_{lh} and ω_{ki} increase by rk ΔW_j, and no other interval multiplicity changes. Hence, the interval multiplicities encode the changes produced by these combinatorial moves more elegantly than the rank list does.

We specify two types of small combinatorial moves. Every subspace ϕ_{lkjih} in Θ_O has indices satisfying $k + 1 \geq j \geq i$. A *connecting move* is a small one-matrix combinatorial move with displacement $\Delta\theta \in \phi_{lkjih}$ in the special case where $k + 1 = j = i$. In a connecting move, ω_{ki} does not exist (as $k < i$) and only three interval multiplicities change. Figure 9 illustrates two examples of connecting moves and offers an intuitive way to interpret them: a connecting move deletes rk ΔW_j copies of the interval $[h, k]$ and rk ΔW_j copies of the interval $[i, l]$, and replaces them with rk ΔW_j copies of the interval $[h, l]$. We think of this as connecting the intervals $[h, k]$ and $[i, l]$ together with an added edge $[j - 1, j] = [k, i]$ to create an interval $[h, l]$; hence
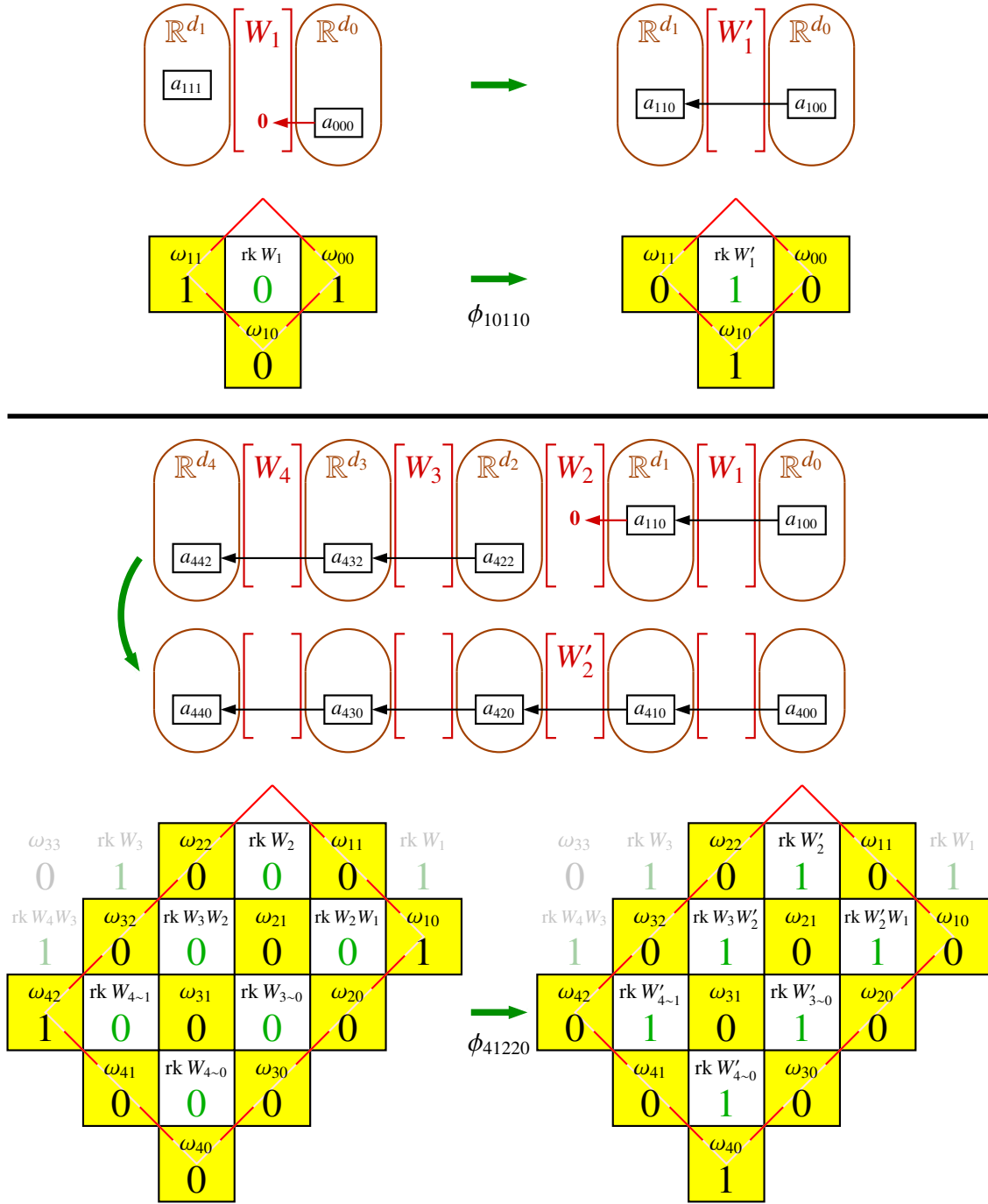
Figure 9: Two examples of connecting moves. The top example is the simplest example possible: W_1 has been perturbed to increase its rank by one. In the bottom example, W_2 has been perturbed. In both examples, the perturbation of W_j causes two intervals $[h, j-1]$ and $[j, l]$ to be replaced by a single 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.

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.) The *rank* of a connecting move is rk ΔW_j.

A *swapping move* is a small one-matrix combinatorial move with displacement $\Delta\theta \in \phi_{lkjih}$ in the case where $k \geq i$. A swapping move changes four interval multiplicities. Figure 10 illustrates two examples of swapping moves. A swapping move splices rk ΔW_j copies of the interval $[h, k]$ with rk ΔW_j copies of the interval $[i, l]$, thereby replacing them with rk ΔW_j copies of the interval $[h, l]$ (which is longer than both of the replaced intervals) and rk ΔW_j copies of the interval $[i, k]$ (which is shorter than both). In effect, the interval endpoints are swapped. The *rank* of a swapping move is rk ΔW_j.

The ideas of connecting and swapping moves, along with Figures 9 and 10, expose much intuition about how the strata are connected to each other. A small move whose displacement comes from a subspace in Θ_O gives us some simple ways that an infinitesimal perturbation of a point in weight space can move you from one stratum to another stratum (always of higher dimension). However, there also exist small moves (not chosen from a single subspace in Θ_O) that are equivalent to a sequence of connecting and swapping moves. In Section 7.8, we will see that every small combinatorial move is equivalent to a sequence of connecting and swapping moves.

We define the following sets of one-matrix subspaces that correspond to connecting moves, swapping moves, and the union of both (small combinatorial moves). We also define a one-matrix subspace that corresponds to moves that stay on the same stratum, omitting all combinatorial moves.

$$
\begin{aligned}
\Theta_O^{\text{conn}} &= \{\phi_{lkjih} \in \Theta_O : l \geq k + 1 = j = i > h\} = \{\phi_{l,j-1,j,j,h} \neq \{\mathbf{0}\} : L \geq l \geq j > h \geq 0\}, \\
\Theta_O^{\text{swap}} &= \{\phi_{lkjih} \in \Theta_O : l > k \geq i > h\} = \{\phi_{lkjih} \neq \{\mathbf{0}\} : L \geq l > k \geq i > h \geq 0 \text{ and } k + 1 \geq j \geq i\}, \\
\Theta_O^{\text{comb}} &= \Theta_O^{\text{conn}} \cup \Theta_O^{\text{swap}} \\
&= \{\phi_{lkjih} \in \Theta_O : l > k \text{ and } i > h\} = \{\phi_{lkjih} \neq \{\mathbf{0}\} : L \geq l \geq k + 1 \geq j \geq i > h \geq 0\}, \\
\Theta_O^{\text{stratum}} &= \Theta_O^{\text{fiber}} \setminus \Theta_O^{\text{comb}} = \Theta_O \setminus \Theta_O^{L0} \setminus \Theta_O^{\text{comb}} = \{\phi_{lkjih} \in \Theta_O : (L > l \text{ or } h > 0) \text{ and } (l \leq k \text{ or } i \leq h)\}.
\end{aligned}
$$

For example, in the two-matrix case ($L = 2$),

$$
\begin{aligned}
\Theta_O^{\text{conn}} &= \{\phi_{10110}, \phi_{20110}, \phi_{21220}, \phi_{21221}\} \setminus \{\{\mathbf{0}\}\}, \\
\Theta_O^{\text{swap}} &= \{\phi_{21110}, \phi_{21210}\} \setminus \{\{\mathbf{0}\}\}, \\
\Theta_O^{\text{comb}} &= \{\phi_{10110}, \phi_{20110}, \phi_{21110}, \phi_{21210}, \phi_{21220}, \phi_{21221}\} \setminus \{\{\mathbf{0}\}\}, \\
\Theta_O^{\text{stratum}} &= \{\phi_{10100}, \phi_{11100}, \phi_{11110}, \phi_{12100}, \phi_{12110}, \phi_{21201}, \phi_{21211}, \phi_{22201}, \phi_{22211}, \phi_{22221}\} \setminus \{\{\mathbf{0}\}\}.
\end{aligned}
$$

7.5 Abstract Moves

Each type of connecting or swapping move has an effect on the rank list (and interval multiplicities) that depends solely on the indices h, i, k, and l and the rank of ΔW_j. This motivates the idea of an *abstract move* that maps one rank list to another rank list, divorced entirely from any geometry. A *rank-c abstract connecting move* takes a valid rank list $\underline{r}$ and an index tuple (l, k, i, h) satisfying $L \geq l \geq k + 1 = i > h \geq 0$, and yields the modified rank list $\underline{s}$ produced by increasing ω_{lh} by c and decreasing ω_{li} and ω_{kh} by c. A *rank-c abstract swapping move* takes a valid rank list $\underline{r}$ and an index tuple (l, k, i, h) satisfying $L \geq l > k \geq i > h \geq 0$, and yields the modified rank list $\underline{s}$ produced by increasing ω_{lh} and ω_{ki} by c and decreasing ω_{li} and ω_{kh} by c. We must have $\omega_{li} \geq c$ and $\omega_{kh} \geq c$ prior to either type of move, so that the move produces a valid rank list.

In Section 7.8, we will see that for valid rank lists satisfying $\underline{r} < \underline{s}$, there exists a sequence of rank-1 abstract connecting and swapping moves that proceed from $\underline{r}$ to $\underline{s}$. For that reason, in Section 3.1 we introduced the convention that each edge of the stratum dag represents a rank-1 abstract connecting or swapping move.
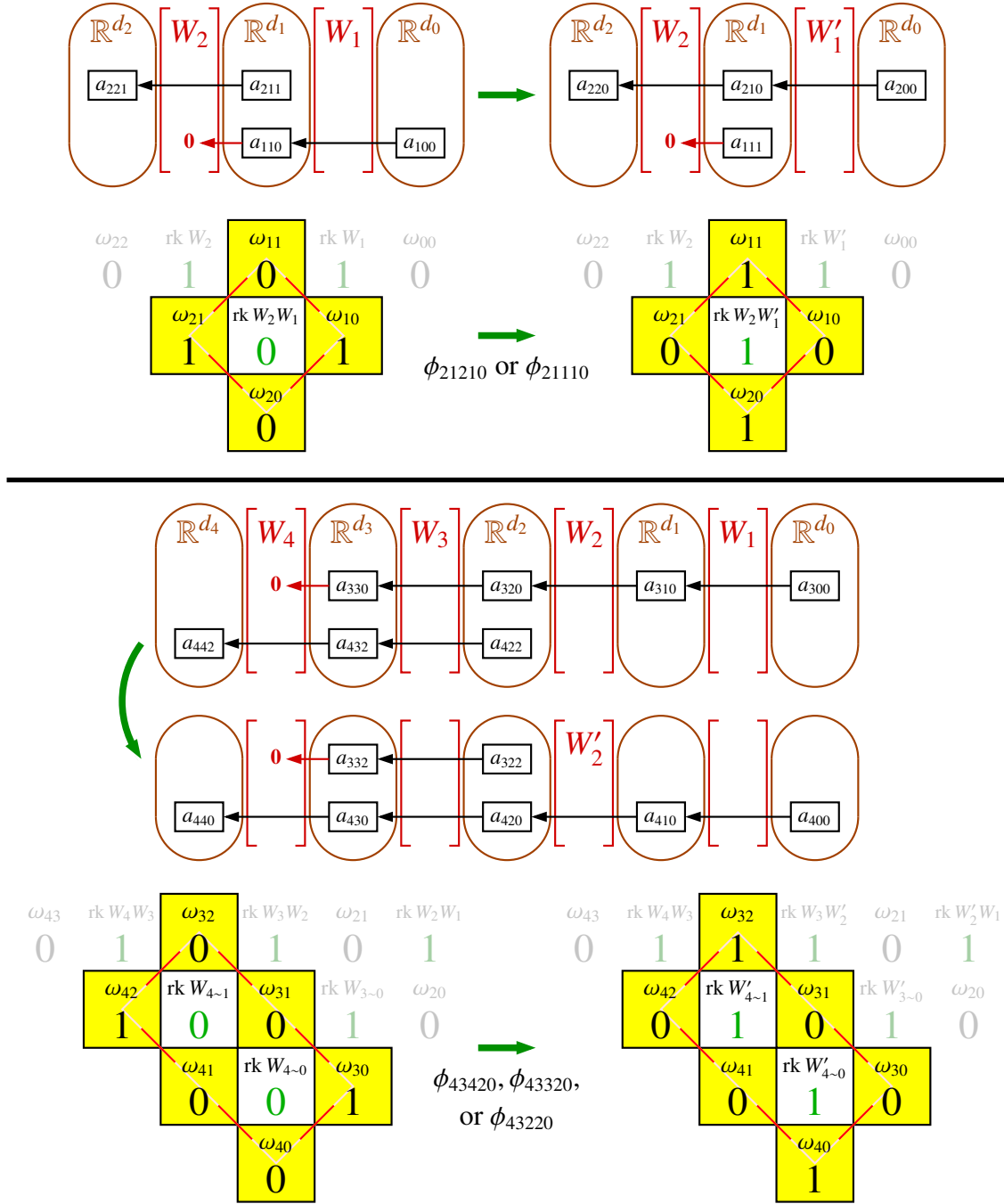
Figure 10: Two examples of swapping moves. In the top example—the simplest example possible—either W_1 or W_2 may be perturbed to cause the move. In the bottom example, any of of W_2, W_3, or W_4 may have been perturbed. Two intervals $[h, k]$ and $[i, l]$ are replaced by an interval $[h, l]$, longer 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.

In Section 3.1 we noted that sometimes a stratum dag contains a directed edge $(S_{\underline{r}}, S_{\underline{s}})$ despite also containing the edges $(S_{\underline{r}}, S_{\underline{t}})$ and $(S_{\underline{t}}, S_{\underline{s}})$—that is, the dag is not a Hasse diagram. An example illustrated in Figure 11 explains why. Consider a stratum dag that includes the four strata whose basis flow diagrams are depicted. (Suppose that there are four layers of edges with $W_4 = 0$ but the figure shows only the first three; thus all four strata are on the same fiber.) From the stratum $S_{\underline{r}}$, depicted at upper left, there are rank-1 swapping moves that move directly onto the stratum $S_{\underline{s}}$ at lower right. But it is also possible to move from $S_{\underline{r}}$ to $S_{\underline{s}}$ through a sequence of two rank-1 swapping moves, passing through $S_{\underline{t}}$ or $S_{\underline{u}}$ along the way. Thus the edge $(S_{\underline{r}}, S_{\underline{s}})$ is redundant for the purpose of diagramming a partial order of strata. Nevertheless, we include $(S_{\underline{r}}, S_{\underline{s}})$ in the dag because the one-matrix subspaces ϕ_{31210} and ϕ_{31110} (as they are defined at a point $\theta \in S_{\underline{r}}$) represent degrees of freedom of direct motion from $S_{\underline{r}}$ into $S_{\underline{s}}$ that are linearly independent of the one-matrix subspaces that represent moves into $S_{\underline{t}}$ or $S_{\underline{u}}$ (ϕ_{32310}, ϕ_{32210}, ϕ_{32110}, ϕ_{21210}, and ϕ_{21110}). These degrees of freedom of motion into $S_{\underline{s}}$ are not represented by the dag's indirect paths from $S_{\underline{r}}$ to $S_{\underline{s}}$, but they are important for understanding how the strata meet each other geometrically.

7.6 One-Matrix Moves on a Stratum

Given a starting point θ on a stratum S, the significance of $\Theta_O^{\text{stratum}}$ is that for any subspace $\phi_{lkjih} \in \Theta_O^{\text{stratum}}$, a move with a displacement $\Delta\theta \in \phi_{lkjih}$ stays on $\bar{S}$. (The move stays on the fiber and does not increase any matrix rank.) If it is a small move, it stays on S proper. (It does not increase nor decrease any matrix rank.) We can rephrase the first property to say that $\bar{S}$ includes the affine subspace obtained by translating ϕ_{lkjih} so it passes through θ, which we can write $\{\theta + \Delta\theta : \Delta\theta \in \phi_{lkjih}\}$. This also gives us some information about $T_\theta S$, the space tangent to S at θ, because it implies that $\phi_{lkjih} \subseteq T_\theta S$.

Lemma 25. *Every* $\phi_{lkjih} \in \Theta_O^{\text{stratum}}$ *satisfies* $\{\theta + \Delta\theta : \Delta\theta \in \phi_{lkjih}\} \subseteq \bar{S}$ *and* $\phi_{lkjih} \subseteq T_\theta S$. *Moreover,* span $\Theta_O^{\text{stratum}} \subseteq T_\theta S$.

Proof. Consider a subspace $\phi_{lkjih} \in \Theta_O^{\text{stratum}}$ and a displacement $\Delta\theta \in \phi_{lkjih}$. Recall that $\Theta_O^{\text{stratum}} = \Theta_O^{\text{fiber}} \setminus \Theta_O^{\text{comb}}$. As $\phi_{lkjih} \in \Theta_O^{\text{fiber}}$, by Corollary 21, $\mu(\theta + \Delta\theta) = W$. Hence the translated subspace $\{\theta + \Delta\theta : \Delta\theta \in \phi_{lkjih}\}$ is a subset of the fiber $\mu^{-1}(W)$.

Consider a parametrized point $\theta' = \theta + \epsilon\Delta\theta$ with $\epsilon \in \mathbb{R}$, describing a line though θ and $\theta + \Delta\theta$, and the subsequence matrices $W'_{y\sim x}$ derived from θ'. As $\phi_{lkjih} \notin \Theta_O^{\text{comb}}$, either $l \leq k$ or $i \leq h$. By Lemma 24, rk $W'_{y\sim x} \leq$ rk $W_{y\sim x}$ for every subsequence matrix. Let $\underline{r}$ be the rank list of θ and let $\underline{r}'$ be the rank list of θ'; thus $\underline{r}' \leq \underline{r}$. We will show that on the line though θ and $\theta + \Delta\theta$, there are only finitely many points where the inequality is strict—that is, where $\underline{r}' < \underline{r}$—and thus almost all the points on the line are in the stratum S, and thus the line is a subset of $\bar{S}$.

For a specific choice of y and x satisfying $L \geq y \geq x \geq 0$, the rank of the subsequence matrix $W_{y\sim x}$ is $r_{y\sim x}$, so there is some $r_{y\sim x} \times r_{y\sim x}$ minor of $W_{y\sim x}$ that has a nonzero determinant. Consider the same minor of $W'_{y\sim x}$; the determinant of that minor is a polynomial function of ϵ that is nonzero at $\epsilon = 0$. Hence it has finitely many zeros (with respect to ϵ), hence there are only finitely many points on the line though θ and $\theta + \Delta\theta$ where rk $W'_{y\sim x} <$ rk $W_{y\sim x}$. At all other points on the line, rk $W'_{y\sim x} =$ rk $W_{y\sim x}$. As the number of subsequence matrices is finite, there are only finitely many points on the line where $\underline{r}' < \underline{r}$, as claimed; the line is a subset of $\bar{S}$. This reasoning holds for every displacement $\Delta\theta \in \phi_{lkjih}$, so $\{\theta + \Delta\theta : \Delta\theta \in \phi_{lkjih}\} \subseteq \bar{S}$.

By Theorem 16, S is a smooth manifold, hence a tangent space $T_\theta S$ exists. The line though θ and $\theta + \Delta\theta$ described above is tangent to S at θ, so the parallel line though the origin and $\Delta\theta$ is a subset of $T_\theta S$. This reasoning holds for every displacement $\Delta\theta \in \phi_{lkjih}$, so $\phi_{lkjih} \subseteq T_\theta S$.

It follows that span $\Theta_O^{\text{stratum}} \subseteq T_\theta S$ because $T_\theta S$ is a subspace and $\phi_{lkjih} \subseteq T_\theta S$ for every $\phi_{lkjih} \in \Theta_O^{\text{stratum}}$. $\square$
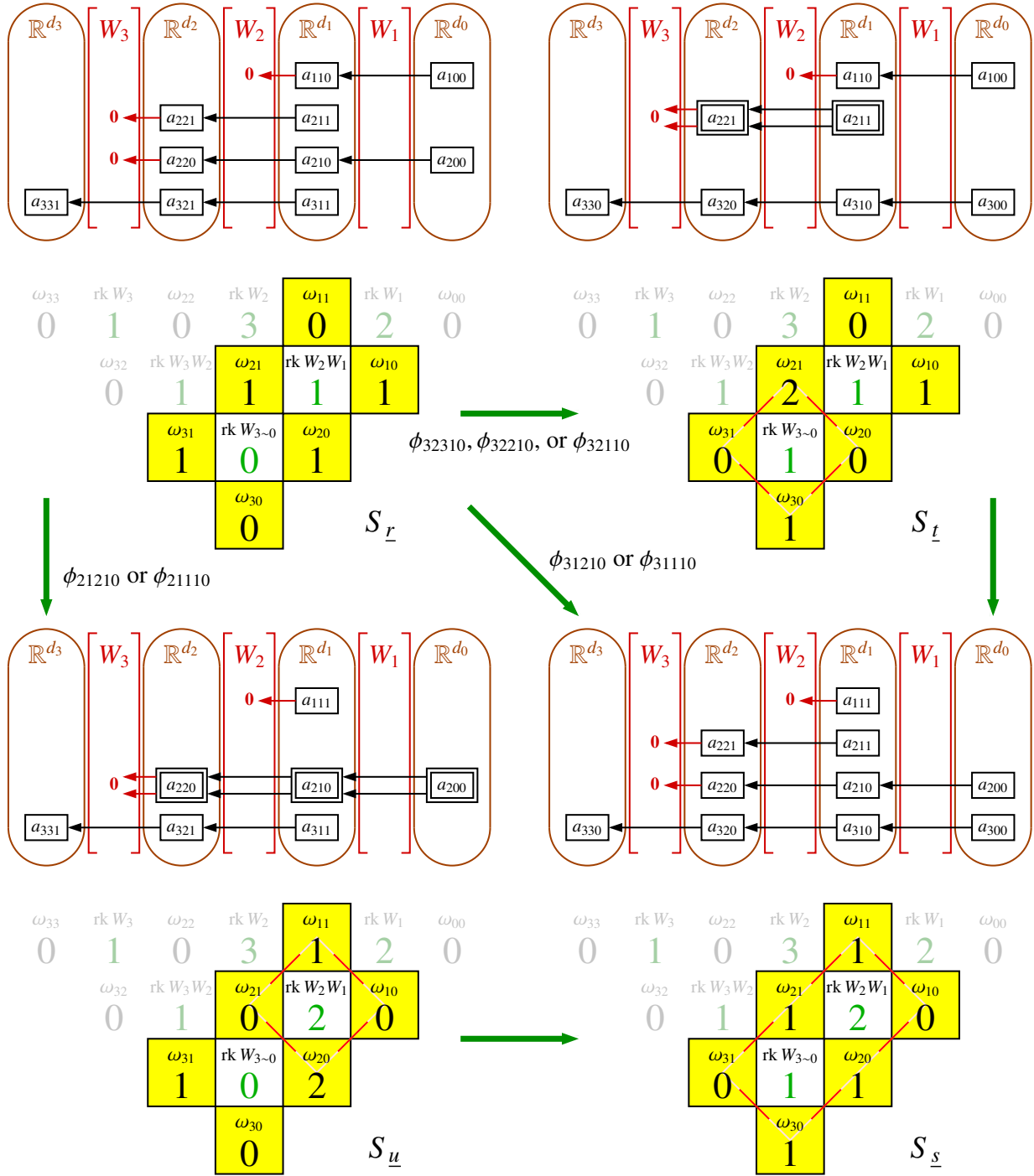
Figure 11: Basis flow diagrams for four strata in a stratum dag. Green arrows are edges of the dag; in this example they all representing swapping moves. This dag includes the directed edge $(S_{\underline{r}}, S_{\underline{s}})$, representing swapping moves whose displacements come from the subspaces ϕ_{31210} or ϕ_{31110}, even though there are directed paths from $S_{\underline{r}}$ to $S_{\underline{s}}$ passing through $S_{\underline{t}}$ or $S_{\underline{u}}$. Hence the dag is not a Hasse diagram.

Knowing that $\operatorname{span}\Theta_{\mathrm{O}}^{\text{stratum}} \subseteq T_\theta S$, we are motivated to write an explicit expression for $\operatorname{span}\Theta_{\mathrm{O}}^{\text{stratum}}$. We will use it to derive an expression for $T_\theta S$ in Section 8.5.

Lemma 26.

$$
\begin{aligned}
\operatorname{span}\Theta_{\mathrm{O}}^{\text{stratum}} \;=\; & \Big\{ (\Delta W_L, \Delta W_{L-1}, \ldots, \Delta W_1) : \\[4pt]
& \Delta W_j \in \sum_{h=1}^{j-1} A_{Ljh} \otimes B_{j-1,j-1,h} + A_{L-1,j,0} \otimes B_{j-1,j-1,0} + \\[4pt]
& \sum_{l=j}^{L-1} A_{ljj} \otimes B_{l,j-1,0} + A_{Ljj} \otimes B_{L,j-1,1} \Big\} \tag{7.1} \\[8pt]
\;=\; & \Big\{ (\Delta W_L, \Delta W_{L-1}, \ldots, \Delta W_1) : \\[4pt]
& \Delta W_j \in \sum_{h=1}^{j-1} \operatorname{col} W_{j\sim h} \otimes \operatorname{null} W_{j-1\sim h-1}^\top + (\operatorname{null} W_{L\sim j} \cap \operatorname{col} W_{j\sim 0}) \otimes \mathbb{R}^{d_{j-1}} + \\[4pt]
& \sum_{l=j}^{L-1} \operatorname{null} W_{l+1\sim j} \otimes \operatorname{row} W_{l\sim j-1} + \mathbb{R}^{d_j} \otimes (\operatorname{row} W_{L\sim j-1} \cap \operatorname{null} W_{j-1\sim 0}^\top) \Big\}. \tag{7.2}
\end{aligned}
$$

Proof. By the definition of $\Theta_{\mathrm{O}}^{\text{stratum}}$,

$$
\begin{aligned}
\operatorname{span}\Theta_{\mathrm{O}}^{\text{stratum}} \;=\; & \operatorname{span}\{\phi_{lkjih} \in \Theta_{\mathrm{O}} : (L > l \text{ or } h > 0) \text{ and } (l \le k \text{ or } i \le h)\} \\
\;=\; & \Big\{ (\Delta W_L, \Delta W_{L-1}, \ldots, \Delta W_1) : \\
& \Delta W_j \in \operatorname{span}\{o_{lkjih} \in O_j : (L > l \text{ or } h > 0) \text{ and } (l \le k \text{ or } i \le h)\} \Big\}. \tag{7.3}
\end{aligned}
$$

The condition "$(L > l$ or $h > 0)$ and $l \le k$" is equivalent to the condition "$(L > l$ and $l \le k)$ or $(h > 0$ and $k = L)$": the forward direction follows because $l \le k$ and **not** $L > l$ imply that $k = L$, and the reverse direction follows because $k = L$ implies that $l \le k$. Similarly, the condition "$(L > l$ or $h > 0)$ and $i \le h$" is equivalent to the condition "$(h > 0$ and $i \le h)$ or $(L > l$ and $i = 0)$": the forward direction follows because $i \le h$ and **not** $h > 0$ imply $i = 0$, and the reverse direction follows because $i = 0$ implies $i \le h$. Thus we can rewrite part of (7.3) as

$$
\begin{aligned}
& \operatorname{span}\{o_{lkjih} \in O_j : (L > l \text{ or } h > 0) \text{ and } (l \le k \text{ or } i \le h)\} \\
=\; & \operatorname{span}\{o_{lkjih} \in O_j : (L > l \text{ or } h > 0) \text{ and } l \le k\} + \operatorname{span}\{o_{lkjih} \in O_j : (L > l \text{ or } h > 0) \text{ and } i \le h\} \\
=\; & \operatorname{span}\{o_{lkjih} \in O_j : L > l \text{ and } l \le k\} + \operatorname{span}\{o_{lkjih} \in O_j : h > 0 \text{ and } k = L\} + \\
& \operatorname{span}\{o_{lkjih} \in O_j : h > 0 \text{ and } i \le h\} + \operatorname{span}\{o_{lkjih} \in O_j : L > l \text{ and } i = 0\}. \tag{7.4}
\end{aligned}
$$

To link the spans in (7.4) with the tensor products in (7.1), we use Lemma 4 to write

$$
\sum_{h=1}^{j-1} A_{Ljh} \otimes B_{j-1,j-1,h} = \sum_{h=1}^{j-1}\sum_{l=j}^{L}\sum_{i=0}^{h} a_{lji} \otimes \sum_{k=j-1}^{L}\sum_{h'=h}^{j-1} b_{k,j-1,h'} = \sum_{h=1}^{j-1}\sum_{l=j}^{L}\sum_{i=0}^{h}\sum_{k=j-1}^{L}\sum_{h'=h}^{j-1} o_{lkjih'}
$$

$$
= \operatorname{span}\{o_{lkjih} \in O_j : h > 0 \text{ and } i \leq h\}, \tag{7.5}
$$

$$
A_{L-1,j,0} \otimes B_{j-1,j-1,0} = \sum_{l=j}^{L-1} a_{lj0} \otimes \sum_{k=j-1}^{L}\sum_{h=0}^{j-1} b_{k,j-1,h} = \sum_{l=j}^{L-1}\sum_{k=j-1}^{L}\sum_{h=0}^{j-1} o_{lkj0h}
$$

$$
= \operatorname{span}\{o_{lkjih} \in O_j : L > l \text{ and } i = 0\}, \tag{7.6}
$$

$$
\sum_{l=j}^{L-1} A_{ljj} \otimes B_{l,j-1,0} = \sum_{l=j}^{L-1}\sum_{l'=j}^{l}\sum_{i=0}^{j} a_{l'ji} \otimes \sum_{k=l}^{L}\sum_{h=0}^{j-1} b_{k,j-1,h} = \sum_{l=j}^{L-1}\sum_{l'=j}^{l}\sum_{i=0}^{j}\sum_{k=l}^{L}\sum_{h=0}^{j-1} o_{l'kjih}
$$

$$
= \operatorname{span}\{o_{lkjih} \in O_j : L > l \text{ and } l \leq k\}, \quad \text{ and} \tag{7.7}
$$

$$
A_{Ljj} \otimes B_{L,j-1,1} = \sum_{l=j}^{L}\sum_{i=0}^{j} a_{lji} \otimes \sum_{h=1}^{j-1} b_{L,j-1,h} = \sum_{l=j}^{L}\sum_{i=0}^{j}\sum_{h=1}^{j-1} o_{lLjih}
$$

$$
= \operatorname{span}\{o_{lkjih} \in O_j : h > 0 \text{ and } k = L\}. \tag{7.8}
$$

The claim (7.1) follows from (7.3)–(7.8). The claim (7.2) follows from (7.1) and the flow subspace definitions (4.4), (4.5). □

For example, in the two-matrix case ($L = 2$),

$$
\operatorname{span}\Theta_{\mathrm{O}}^{\mathrm{stratum}} = \Big\{(\Delta W_2, \Delta W_1) :
$$
$$
\Delta W_2 \in \operatorname{col} W_2 \otimes \operatorname{null} W_1^{\top} + \mathbb{R}^{d_2} \otimes (\operatorname{row} W_2 \cap \operatorname{null} W_1^{\top}),
$$
$$
\Delta W_1 \in (\operatorname{null} W_2 \cap \operatorname{col} W_1) \otimes \mathbb{R}^{d_0} + \operatorname{null} W_2 \otimes \operatorname{row} W_1\Big\}.
$$

The formula (7.2) is the most explicit and concise expression we know how to write for $\operatorname{span}\Theta_{\mathrm{O}}^{\mathrm{stratum}}$, but it does not readily reveal the dimension of $\operatorname{span}\Theta_{\mathrm{O}}^{\mathrm{stratum}}$ nor a basis that spans it, because the formula uses a vector sum of subspaces that are very far from being linearly independent. The easiest way to find a basis is to explicitly compute $\Theta_{\mathrm{O}}^{\mathrm{stratum}}$. We derive the dimension of $\operatorname{span}\Theta_{\mathrm{O}}^{\mathrm{stratum}}$ in Section 8.7.

7.7 Some Intuition for the One-Matrix Subspaces

We can gain some intuition about the one-matrix prebasis by inspecting the one-matrix subspaces for the canonical weight vector $\tilde{\theta} = (\tilde{I}_L, \tilde{I}_{L-1}, \ldots, \tilde{I}_1)$ defined in Section 4.6. At the canonical weight vector, the one-matrix subspace ϕ_{lkjih} is defined by permitting a single block of $\Delta\tilde{I}_j$ to vary arbitrarily, namely, the block whose rows are associated with the interval $[i, l]$ and whose columns are associated with the interval $[h, k]$, while setting all the other components of the displacement $\Delta\theta$ to zero.

To clarify this idea, see Figure 12, which reprises part of Figure 8 with some background colors added. The one-matrix subspace ϕ_{22100} is the set of displacements $\Delta\theta$ that can be obtained by arbitrarily varying the 2×2 block of $\Delta\tilde{I}_1$ whose rows and columns are associated with the interval $[0, 2]$ (labeled "20" in the figure), while setting $\Delta\tilde{I}_4$, $\Delta\tilde{I}_3$, $\Delta\tilde{I}_2$, and all the other components of $\Delta\tilde{I}_1$ to zero. At $\tilde{\theta}$, every one-matrix subspace is *axis-aligned*, meaning that it is spanned by a subset of the coordinate axes of $\mathbb{R}^{d_\theta}$.
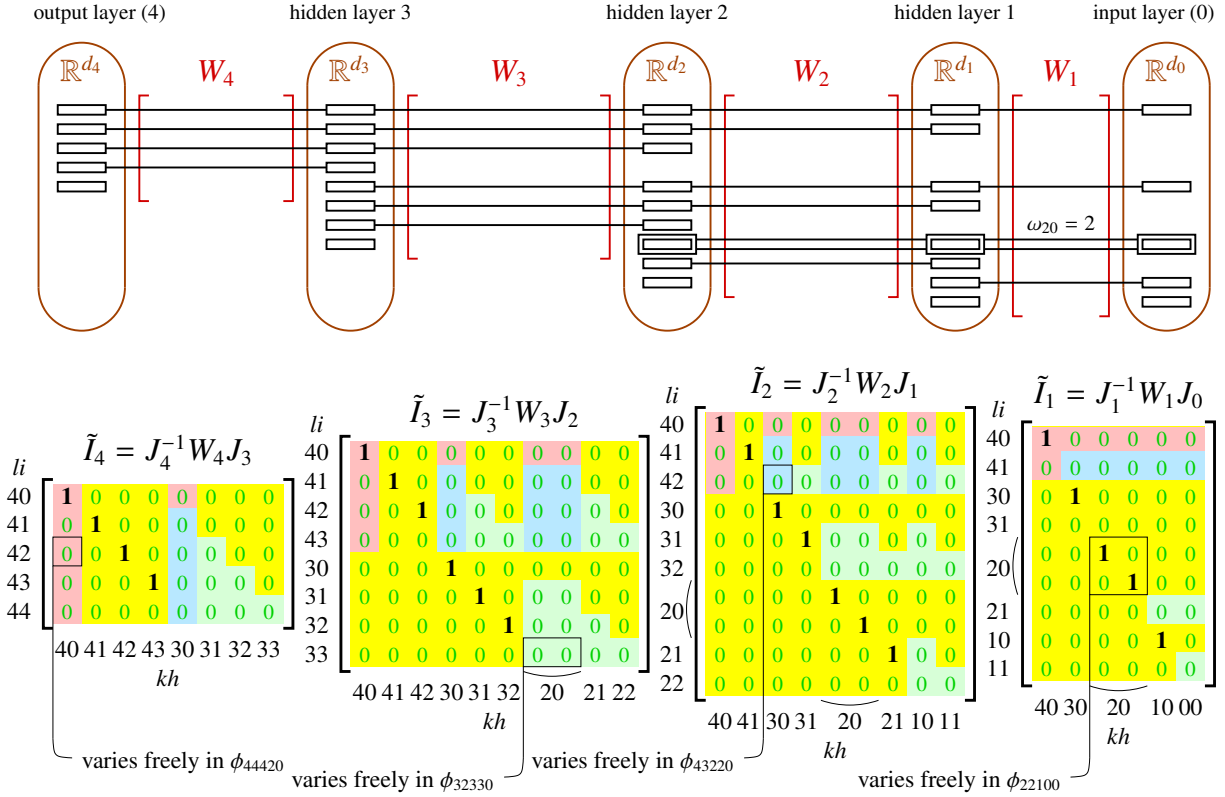
Figure 12: The canonical weight vector $\tilde{\theta} = (\tilde{I}_4, \tilde{I}_3, \tilde{I}_2, \tilde{I}_1)$ when every $\omega_{ki} = 1$ except $\omega_{20} = 2$. Yellow and green backgrounds indicate moves that stay on the fiber (moves with displacements from subspaces in Θ_O^{fiber} such as ϕ_{22100} and ϕ_{32330}), whereas pink and blue backgrounds indicate moves off the fiber ($l = 4$ and $h = 0$, moves from subspaces in Θ_O^{L0} such as ϕ_{44420} and ϕ_{43220}). Green and blue backgrounds indicate combinatorial moves that increase the rank of some subsequence matrix ($l > k$ and $i > h$, moves from subspaces in Θ_O^{comb} such as ϕ_{32330} and ϕ_{43220}), whereas yellow and pink backgrounds indicate moves that do not increase any rank. Hence, yellow indicates moves that stay on the closure of the stratum (moves from subspaces in $\Theta_O^{\text{stratum}}$ such as ϕ_{22100}). Green indicates combinatorial moves to a different stratum that stay on the fiber. Pink indicates moves off the fiber in which no subsequence matrix's rank increases. Blue indicates combinatorial moves off the fiber.

In the figure, we have given each component of each factor matrix a background color to distinguish the effects of a one-matrix move. In any single factor matrix $\tilde{I}_j$, components with a yellow background can be changed without changing the product $\tilde{I}_4\tilde{I}_3\tilde{I}_2\tilde{I}_1$ nor increasing the rank of any subsequence matrix. That is, a change to a yellow component (or a yellow block with common indices) is a move whose displacement is in a subspace in $\Theta_O^{\text{stratum}}$, such as ϕ_{22100}. This move stays on the fiber by Corollary 21 and, by Lemma 25, it also stays on the stratum's closure. (If the move is small, it stays on the stratum proper. If the move is not small, it might not stay on the stratum, because some subsequence matrix's rank might decrease; but no subsequence matrix's rank can increase.)

A green background marks components whose change increases the rank of some subsequence matrix but does not change $\tilde{I}_4\tilde{I}_3\tilde{I}_2\tilde{I}_1$—hence staying on the fiber but moving to a different stratum. That is, a change to a green component is a move whose displacement is in a subspace in $\Theta_O^{\text{fiber}} \cap \Theta_O^{\text{comb}}$, such as ϕ_{32330}. Pink marks components whose change changes $\tilde{I}_4\tilde{I}_3\tilde{I}_2\tilde{I}_1$ but increases no rank—hence moving off the fiber. A change to a pink component is a move whose displacement is in a subspace in Θ_O^{L0} but not Θ_O^{comb}, such as ϕ_{44420}.

Blue marks components whose change changes $\tilde{I}_4 \tilde{I}_3 \tilde{I}_2 \tilde{I}_1$ and increases its rank—combinatorial moves off the fiber. A change to a blue component is a move whose displacement is in a subspace in $\Theta_O^{L0} \cap \Theta_O^{comb}$, such as ϕ_{43220}.

This intuition can be extended from the canonical weight vector $\tilde{\theta}$ for a rank list $\underline{r}$ to any other weight vector θ that also has rank list $\underline{r}$. Let $W = \mu(\theta)$ and recall from Section 5.2 the linear function $\eta : (X_L, X_{L-1}, \ldots, X_1) \mapsto (J_L X_L J_{L-1}^{-1}, J_{L-1} X_{L-1} J_{L-2}^{-1}, \ldots, J_1 X_1 J_0^{-1})$ that maps the fiber of $\tilde{I}$ to the fiber of W. An invertible linear transformation of weight space preserves tangencies, so the function η maps the one-matrix subspaces at $\tilde{\theta}$ (which are axis-aligned) to the one-matrix subspaces at θ (which are not).

7.8 The Hierarchy of Strata

At the beginning of Section 7, we made the following claim about the stratum interconnections. Now that we have defined connecting and swapping moves, we are adding one extra statement to the list: statement D below is equivalent to saying that there is a directed path from $S_{\underline{r}}$ to $S_{\underline{s}}$ in the stratum dag.

Theorem 27. *Let $\underline{r}$ and $\underline{s}$ be two valid rank lists for the same linear neural network (i.e., $r_{j\sim j} = s_{j\sim j} = d_j$ for all $j \in [0, L]$). Let $\mu^{-1}(W)$ be the fiber of a matrix $W \in \mathbb{R}^{d_L \times d_0}$ whose rank satisfies $\mathrm{rk}\, W = r_{L\sim 0} = s_{L\sim 0}$. Let $S_{\underline{r}}$ and $S_{\underline{s}}$ be the strata with rank lists $\underline{r}$ and $\underline{s}$ in the rank stratification of $\mu^{-1}(W)$, and observe that both strata are nonempty by Lemma 55. Then the following statements are equivalent (imply each other).*

- *A. $S_{\underline{r}} \subseteq \bar{S}_{\underline{s}}$.*

- *B. $S_{\underline{r}} \cap \bar{S}_{\underline{s}} \neq \emptyset$.*

- *C. $\underline{r} \leq \underline{s}$.*

- *D. There exists a sequence of rank-one abstract connecting and swapping moves that proceed from $\underline{r}$ to $\underline{s}$, with all the intermediate rank lists being valid.*

Proof. Given the assumption that $S_{\underline{r}} \neq \emptyset$, it is clear that A implies B. Lemma 17 states that B implies C. The forthcoming Corollary 32 states that C implies D. The forthcoming Lemma 28 states that D implies A. □

Observe that while claims A and B are statements about geometry, claims C and D are purely combinatorial. Our proof that C implies D (Lemma 32) is also purely combinatorial (and it does not use the assumption that $S_{\underline{r}}$ and $S_{\underline{s}}$ are nonempty). The fact that C implies D is by far the hardest of the four implications to prove; for many months we did not know if it was true. The fact that B implies A means that our stratifications satisfy the frontier condition (defined in Section 5.2).

We now prove that D implies A. The proof is more opaque than we would like, but the essence of the proof is that every abstract connecting or swapping move can be instantiated as an actual move on the geometry of the fiber—a move that can be arbitrarily small, but not merely an abstract or infinitesimal move.

Lemma 28. *Consider a sequence of valid rank lists $\underline{r}^0, \underline{r}^1, \ldots, \underline{r}^z$ such that each successive rank list can be reached from the previous rank list by an abstract connecting or swapping move (of any rank). Let W be a matrix, and suppose that $r_{L\sim 0}^i = \mathrm{rk}\, W$ for all $i \in [0, z]$. Let $S_{\underline{r}^0}$ and $S_{\underline{r}^z}$ be strata in the rank stratification of the fiber $\mu^{-1}(W)$. Then $S_{\underline{r}^0} \subseteq \bar{S}_{\underline{r}^z}$.*

Proof. If $S_{\underline{r}^0} = \emptyset$, the result follows immediately. Otherwise, let θ^0 be any point in $S_{\underline{r}^0}$. For each $m \in [1, z]$, we will identify a point $\theta^m \in S_{\underline{r}^m}$ that can be reached from θ^{m-1} by a (not abstract, not infinitesimal) connecting or swapping move. By induction, we obtain a point $\theta^z \in S_{\underline{r}^z}$ that is as close to θ^0 as we like.

Assume for the sake of induction that there is a point $\theta^{m-1} \in S_{\underline{r}^{m-1}}$. By assumption there is a rank-c abstract move from $\underline{r}^{m-1}$ to $\underline{r}^m$; it is specified by an index tuple (l, k, i, h) and the rank c. The fact that $\underline{r}^m$ is a valid rank list implies that the previous rank list $\underline{r}^{m-1}$ has interval multiplicities $\omega_{li} \geq c$ and $\omega_{kh} \geq c$ (as the move decreases both multiplicities by c). Choose any $j \in [i, k+1]$ and consider the one-matrix subspace ϕ_{lkjih} defined at the point θ^{m-1}. Its dimension is $\dim \phi_{lkjih} = \omega_{li} \omega_{kh} \geq c^2$, and some matrices in o_{lkjih} have rank c. Let $\Delta\theta \in \phi_{lkjih}$ be a displacement such that $\Delta W_j \in o_{lkjih}$ has rank c. If ΔW_j is sufficiently small, then by Lemma 24, $\theta^{m-1} + \Delta\theta$ has rank list $\underline{r}^m$. We set $\theta^m = \theta^{m-1} + \Delta\theta$.

We can choose each successive displacement sufficiently small that each move in the sequence of moves is a "small move," meaning that no move takes us to a point that is in the closure of any stratum whose closure does not contain θ^0. As the final point θ^z lies on $S_{\underline{r}^z}$, it follows that $\theta^0 \in \bar{S}_{\underline{r}^z}$.

This construction can be applied to every point $\theta^0 \in S_{\underline{r}^0}$, so $S_{\underline{r}^0} \subseteq \bar{S}_{\underline{r}^z}$. $\qquad\square$

We devote the rest of this section to proving that C implies D: if $\underline{r} \leq \underline{s}$, there exists a sequence of rank-one abstract connecting and swapping moves that takes us from $\underline{r}$ to $\underline{s}$. This means that we can always arrange the strata in a dag like those illustrated in Figures 2, 3, and 4, such that each directed edge of the dag represents a rank-one connecting or swapping move. Every inclusion of one stratum in the closure of another is represented by a path in this dag (that is, A is equivalent to D).

Given rank lists $\underline{r}$ and $\underline{s}$ with $\underline{r} < \underline{s}$, finding a sequence of abstract moves that takes us from $\underline{r}$ to $\underline{s}$ is an interesting recreational puzzle, which took us three months to solve. Our solution begins with the algorithm FINDLASTMOVE in Figure 13, which finds a rank list $\underline{t}$ such that $\underline{r} \leq \underline{t} < \underline{s}$ and a single rank-one abstract connecting or swapping move takes us from $\underline{t}$ to $\underline{s}$. Building on this step, a simple recursive algorithm, FINDALLMOVES in Figure 13, finds a sequence of abstract connecting and swapping moves that take us from $\underline{r}$ to $\underline{s}$ (computing the sequence in reverse order). The proof of correctness of FINDALLMOVES, and thus the proof that C implies D, follows by induction.

Our algorithms and proofs use differences between the rank lists $\underline{r}$ and $\underline{s}$. Suppose that the interval multiplicities associated with $\underline{r}$ are ω_{ki} and the interval multiplicities associated with $\underline{s}$ are ω_{ki}^s. Let

$$\Delta r_{k \sim i} = s_{k \sim i} - r_{k \sim i} \quad \text{and} \quad \Delta\omega_{ki} = \omega_{ki}^s - \omega_{ki} \quad \text{for all } L \geq k \geq i \geq 0.$$

As we assume that $\underline{r} \leq \underline{s}$, no $\Delta r_{k \sim i}$ is negative. (No such constraint applies to $\Delta\omega_{ki}$.)

For the sake of proving that C implies D, we assume that $r_{L \sim 0} = s_{L \sim 0}$; but our algorithm and the following proofs do *not* require that $r_{L \sim 0} = s_{L \sim 0}$. Perhaps this extra generality will find a use someday. If $\Delta r_{L \sim 0} > 0$, at least one of the moves produced by FINDALLMOVES will move off the fiber, which is why that case is not relevant to the relationships between the strata of a single fiber $\mu^{-1}(W)$.

The following four lemmas derive properties of the algorithms and prove that they always find a valid sequence of rank lists.

Lemma 29. *For indices satisfying $L \geq k \geq i \geq 0$,*

$$\Delta r_{k \sim i} = \sum_{y=k}^{L} \sum_{x=0}^{i} \Delta\omega_{yx} \quad \text{and}$$

$$\Delta\omega_{ki} = \Delta r_{k \sim i} - \Delta r_{k \sim i-1} - \Delta r_{k+1 \sim i} + \Delta r_{k+1 \sim i-1}.$$

FINDLASTMOVE($\underline{r}$, $\underline{s}$)
{ Given valid rank lists $\underline{r}$ and $\underline{s}$ such that $\underline{r} < \underline{s}$, returns a valid rank list $\underline{t}$ such that $\underline{r} \leq \underline{t} < \underline{s}$ and }
{ a single rank-one abstract connecting or swapping move takes us from $\underline{t}$ to $\underline{s}$. }
1   **for all** y and x satisfying $L \geq y \geq x \geq 0$
              { From $\underline{r}$ and $\underline{s}$, compute all the interval multiplicities with (4.10). }
2               $\omega_{yx} \leftarrow r_{y\sim x} - r_{y\sim x-1} - r_{y+1\sim x} + r_{y+1\sim x-1}$ { use the convention that $r_{L+1\sim x} = r_{y\sim -1} = 0$ }
3               $\omega_{yx}^s \leftarrow s_{y\sim x} - s_{y\sim x-1} - s_{y+1\sim x} + s_{y+1\sim x-1}$
4               $\Delta\omega_{yx} \leftarrow \omega_{yx}^s - \omega_{yx}$
5               $\Delta r_{y\sim x} \leftarrow s_{y\sim x} - r_{y\sim x}$ { every $\Delta r_{y\sim x}$ is nonnegative, as $\underline{r} < \underline{s}$ }
               { Initialize $\underline{t}$ to be the same as $\underline{s}$. }
6               $t_{y\sim x} \leftarrow s_{y\sim x}$
7      $[h, l] \leftarrow$ the longest interval such that $\Delta r_{l\sim h} > 1$ (i.e., maximize $l - h$)
8      $i' \leftarrow$ the smallest index in $[h + 1, l]$ such that $i' = l$ or $\Delta\omega_{l-1,i'} > 0$ or $\Delta r_{l\sim i'} = 0$
9   **if** $i' = l$ **or** $\Delta\omega_{l-1,i'} > 0$
10         $i \leftarrow i'; k \leftarrow l - 1$
11   **else**
12         $k \leftarrow$ the greatest index in $[i' - 1, l - 2]$ such that $k = i' - 1$ or $\Delta\omega_{ki} > 0$ for some $i \in [h + 1, i']$
13         $i \leftarrow$ the smallest index in $[h + 1, i']$ such that $k = i - 1$ or $\Delta\omega_{ki} > 0$
         { If $k = i - 1$, we perform a reverse connecting move, equivalent to decrementing ω_{lh}^s by one }
         { and incrementing ω_{li}^s and ω_{kh}^s by one. Otherwise, we perform a reverse swapping move, }
         { equivalent to decrementing ω_{lh}^s and ω_{ki}^s by one and incrementing ω_{li}^s and ω_{kh}^s by one. }
         { Update the rank list $\underline{t}$ to reflect the reverse move. }
14   **for** $y \leftarrow k + 1$ **to** l
15   **for** $x \leftarrow h$ **to** $i - 1$
16            $t_{y\sim x} \leftarrow s_{y\sim x} - 1$
17   **return** $\underline{t}$

FINDALLMOVES($\underline{r}$, $\underline{s}$)
{ Given rank lists $\underline{r}$ and $\underline{s}$ such that $\underline{r} \leq \underline{s}$, finds and writes a sequence of rank lists proceeding from }
{ $\underline{r}$ to $\underline{s}$ by a sequence of rank-one abstract connecting and swapping moves. }
18   **if** $\underline{r} \neq \underline{s}$
19         $\underline{t} \leftarrow$ FINDLASTMOVE($\underline{r}$, $\underline{s}$)
20   FINDALLMOVES($\underline{r}$, $\underline{t}$)
21   Write $\underline{s}$

Figure 13: Algorithm to find a sequence of rank-1 abstract connecting and swapping moves that proceed from a rank list $\underline{r}$ to a rank list $\underline{s}$, assuming that $\underline{r} \leq \underline{s}$. FINDALLMOVES operates recursively, repeatedly invoking FINDLASTMOVE to identify the last move in the sequence.

For indices satisfying $L \geq l > k \geq i > h \geq 0$,

$$\Delta r_{k \sim i} - \Delta r_{k \sim h} - \Delta r_{l \sim i} + \Delta r_{l \sim h} = \sum_{y=k}^{l-1} \sum_{x=h+1}^{i} \Delta \omega_{yx}.$$

Proof. The first identity is obtained from $\Delta r_{k \sim i} = s_{k \sim i} - r_{k \sim i}$ by substituting (4.2) for both $r_{k \sim i}$ and $s_{k \sim i}$, then substituting $\omega_{yx}^s - \omega_{yx} = \Delta \omega_{yx}$. The second and third identities follow directly from the first by substitution. Alternatively, the second identity follows from $\Delta \omega_{ki} = \omega_{ki}^s - \omega_{ki}$ by substituting (4.10) for both ω_{ki} and ω_{ki}^s. □

Lemma 30. *Let $\underline{r}$ and $\underline{s}$ be two valid rank lists for the same linear neural network (i.e., $r_{j \sim j} = s_{j \sim j} = d_j$ for all $j \in [0, L]$) such that $\underline{r} < \underline{s}$. Let h, i, k, and l be the indices chosen by Lines 7–13 of* FINDLASTMOVE. *Then*

- *$\Delta r_{y \sim x} \geq 1$ for all $y \in [k + 1, l]$ and $x \in [h, i - 1]$,*
- *$\omega_{lh}^s \geq 1$, and*
- *one of these two hold: $k = i - 1$ or $\omega_{ki}^s \geq 1$.*

Proof. Line 7 of FINDLASTMOVE chooses $[h, l]$ to be the longest interval such that $\Delta r_{l \sim h} \geq 1$. (As $\underline{r} < \underline{s}$, there is at least one choice of h and l for which $\Delta r_{l \sim h} > 0$ and $l > h$.) By Lemma 29, $\Delta \omega_{lh} = \Delta r_{l \sim h} - \Delta r_{l \sim h-1} - \Delta r_{l+1 \sim h} + \Delta r_{l+1 \sim h-1}$. The last three terms on the right-hand side represent longer intervals, so $\Delta \omega_{lh} = \Delta r_{l \sim h} \geq 1$. As $\Delta \omega_{lh} = \omega_{lh}^s - \omega_{lh}$ and ω_{lh} is always nonnegative, $\omega_{lh}^s \geq 1$, verifying the second claim.

Line 8 chooses i' to be the smallest index in $[h + 1, l]$ such that $i' = l$, $\Delta \omega_{l-1, i'} > 0$, or $\Delta r_{l \sim i'} = 0$. Hence $\Delta \omega_{l-1, x} \leq 0$ and $\Delta r_{l \sim x} \geq 1$ for all $x \in [h + 1, i' - 1]$. (Otherwise, a smaller i' would have been chosen. Note that the statement is vacuously true if $i' = h + 1$.) As we have already shown that $\Delta r_{l \sim h} \geq 1$, we have $\Delta r_{l \sim x} \geq 1$ for all $x \in [h, i' - 1]$.

If $i' = l$ or $\Delta \omega_{l-1, i'} > 0$, then Line 10 sets $i = i'$ and $k = l - 1$, and all the lemma's claims hold. The lemma's first claim, that $\Delta r_{y \sim x} \geq 1$ for all $y \in [k + 1, l]$ and $x \in [h, i - 1]$, holds because as we have seen, $\Delta r_{l \sim x} \geq 1$ for all $x \in [h, i' - 1]$. The lemma's third claim holds because if $i' = l$ then $k = i - 1$; whereas if $\Delta \omega_{l-1, i'} > 0$, then $\Delta \omega_{ki} \geq 1$ and thus $\omega_{ki}^s \geq 1$.

The case remains where neither $i' = l$ nor $\Delta \omega_{l-1, i'} > 0$. In that case, $\Delta \omega_{l-1, i'} \leq 0$, thus $\Delta \omega_{l-1, x} \leq 0$ for all $x \in [h + 1, i']$, and Line 8 has chosen i' such that $\Delta r_{l \sim i'} = 0$. Line 12 chooses k to be the greatest index in $[i' - 1, l - 2]$ such that $k = i' - 1$ or $\Delta \omega_{ki} > 0$ for some $i \in [h + 1, i']$. Hence $\Delta \omega_{yx} \leq 0$ for all $y \in [k + 1, l - 1]$ and $x \in [h + 1, i']$. (Otherwise, a greater k would have been chosen.)

We now show that $\Delta r_{y \sim x} \geq 1$ for all $y \in [k + 1, l]$ and $x \in [h, i' - 1]$. We have already shown that $\Delta r_{l \sim x} \geq 1$ for all $x \in [h, i' - 1]$. It remains to show it for each $\Delta r_{y \sim x}$ for $y \in [k + 1, l - 1]$ and $x \in [h, i' - 1]$. By Lemma 29,

$$\Delta r_{y \sim i'} - \Delta r_{y \sim x} - \Delta r_{l \sim i'} + \Delta r_{l \sim x} = \sum_{q=y}^{l-1} \sum_{p=x+1}^{i'} \Delta \omega_{qp} \leq 0.$$

By assumption, $\underline{r} < \underline{s}$ and thus $\Delta r_{y \sim i'} \geq 0$. Recall that $\Delta r_{l \sim i'} = 0$ and $\Delta r_{l \sim x} \geq 1$, so

$$\Delta r_{y \sim x} \geq \Delta r_{y \sim i'} - \Delta r_{l \sim i'} + \Delta r_{l \sim x} \geq 0 - 0 + 1 = 1.$$

Therefore, $\Delta r_{y \sim x} \geq 1$ for all $y \in [k + 1, l]$ and $x \in [h, i' - 1]$ as claimed. As the algorithm chooses i from $[h + 1, i']$, this confirms the lemma's first claim, that $\Delta r_{y \sim x} \geq 1$ for all $y \in [k + 1, l]$ and $x \in [h, i - 1]$.

The lemma's third claim follows because Lines 12 and 13 explicitly choose k and i such that $k = i - 1$ or $\Delta \omega_{ki} > 0$; in the latter case, $\omega_{ki}^s \geq 1$. □

Lemma 31. *Let $\underline{r}$ and $\underline{s}$ be two valid rank lists for the same linear neural network (i.e., $r_{j\sim j} = s_{j\sim j} = d_j$ for all $j \in [0, L]$) such that $\underline{r} < \underline{s}$. Then the rank list $\underline{t}$ written by* FindAllMoves *is a valid rank list that satisfies $\underline{r} \le \underline{t} < \underline{s}$, and a single rank-1 abstract connecting or swapping move changes $\underline{t}$ to $\underline{s}$.*

Proof. Inspection of FindLastMove shows that the indices chosen by Lines 7–13 always satisfy $l > k$, $k + 1 \ge i$, and $i > h$. Line 6 of FindLastMove initializes $\underline{t}$ to be the same rank list as $\underline{s}$, then Lines 14-16 reduce the value of $t_{y\sim x}$ by one for every $y \in [k + 1, l]$ and $x \in [h, i - 1]$. At least one rank is reduced, which confirms that $\underline{t} < \underline{s}$. By Lemma 30, $\Delta r_{y\sim x} = s_{y\sim x} - r_{y\sim x} \ge 1$ for every $y \in [k + 1, l]$ and $x \in [h, i - 1]$, so $t_{y\sim x} \ge r_{y\sim x}$ for all y and x satisfying $L \ge y \ge x \ge 0$, which confirms that $\underline{r} \le \underline{t}$.

If $k + 1 = i$, then a transition from $\underline{t}$ to $\underline{s}$ is a rank-1 abstract connecting move that replaces one copy of $[h, k]$ and one copy of $[i, l]$ with one copy of $[h, l]$, whereas if $k \ge i$, a transition from $\underline{t}$ to $\underline{s}$ is a rank-1 abstract swapping move that replaces one copy of $[h, k]$ and one copy of $[i, l]$ with one copy of $[h, l]$ and one copy of $[i, k]$. If the algorithm were explicitly computing the interval multiplicities ω^t_{yx} associated with $\underline{t}$, we would have $\omega^t_{li} = \omega^s_{li} + 1$, $\omega^t_{kh} = \omega^s_{kh} + 1$, $\omega^t_{lh} = \omega^s_{lh} - 1$, and if $k \ge i$, $\omega^t_{ki} = \omega^s_{ki} - 1$. All the other interval multiplicities associated with $\underline{t}$ are the same as those associated with $\underline{s}$.

To show that $\underline{t}$ is a valid rank list, we show that the multiset of intervals associated with $\underline{t}$ is valid, then apply Lemma 54. Recall that a multiset of intervals represented by interval multiplicities ω^t_{yx}, $L \ge y \ge x \ge 0$, is valid if $d_j = \sum_{y=j}^{L} \sum_{x=0}^{j} \omega^t_{yx}$ for every $j \in [0, L]$. It is straightforward to see that if the multiplicities ω^s_{yx} satisfy these identities, then the multiplicities ω^t_{yx} specified above (following a reverse abstract move) satisfy them too. We also require that every ω^t_{yx} is nonnegative—otherwise, the interval multiplicities do not represent a multiset. By Lemma 30, $\omega^s_{lh} \ge 1$ and either $k = i - 1$ or $\omega^s_{ki} \ge 1$, from which it follows that $\omega^t_{lh} \ge 0$ and either $k = i - 1$ or $\omega^t_{ki} \ge 0$. For every other pair of indices y and x, $\omega^t_{yx} \ge \omega^s_{yx} \ge 0$. Hence every ω^t_{yx} is nonnegative, and $\underline{t}$ is a valid rank list by Lemma 54. □

Corollary 32. *Let $\underline{r}$ and $\underline{s}$ be two valid rank lists for the same linear neural network (i.e., $r_{j\sim j} = s_{j\sim j} = d_j$ for all $j \in [0, L]$) such that $\underline{r} \le \underline{s}$. Then there exists a sequence of rank lists that starts with $\underline{r}$ and ends with $\underline{s}$ such that each rank list after $\underline{r}$ can be obtained from the previous rank list in the sequence by a single rank-1 abstract connecting or swapping move. The algorithm* FindAllMoves *in Figure 13 finds such a sequence.*

Proof. The first claim follows from Lemma 31 by a simple induction. The second claim follows because FindAllMoves implements this constructive inductive proof as a recursive algorithm. □

Corollary 32 completes the proof of Theorem 27.

8   The Tangent Space, the Nullspace of d$\mu(\theta)$, and Prebases that Span Them

Consider a fiber $\mu^{-1}(W)$, a stratum S in its rank stratification, and a weight vector $\theta \in S \subseteq \mu^{-1}(W) \subset \mathbb{R}^{d_\theta}$. We wish to identify $T_\theta S$, the space tangent to S at θ. Moreover, we wish to identify $T_\theta \bar{S}'$ (if it exists) for every stratum S' whose closure contains θ. (These strata satisfy $\bar{S}' \supset S$, as our stratification satisfies the frontier condition by Lemma 27.)

We identified the one-matrix subspaces that lie in $T_\theta S$ in Lemma 25 (Section 7.4), but they do not generally suffice to span $T_\theta S$. In this section, we find a complete prebasis for $T_\theta S$ by adding some *two-matrix subspaces* related to two-matrix moves. These subspaces represent some directions along which S is curved. The prebasis will help us to determine the dimension of S (which equals the dimension of $T_\theta S$).

We will also study another subspace, the nullspace of the differential map of $\mu(\theta)$, which we define and explain in Section 8.1. This differential map is written $d\mu(\theta)$ and its nullspace is written $\text{null}\, d\mu(\theta)$. The significance of the differential map is that every smooth path on the fiber leaving θ is necessarily tangent to $\text{null}\, d\mu(\theta)$. However, *not* every direction in $\text{null}\, d\mu(\theta)$ is necessarily tangent to a smooth path on the fiber; only some directions are. But if we take all the vectors tangent to smooth paths on the fiber at θ, their vector sum is $\text{null}\, d\mu(\theta)$. This implies that $T_\theta S \subseteq \text{null}\, d\mu(\theta)$ and, for every stratum S' with $\theta \in \bar{S}'$ such that $T_\theta \bar{S}'$ exists, $T_\theta \bar{S}' \subseteq \text{null}\, d\mu(\theta)$. The dimension of $\text{null}\, d\mu(\theta)$ is the number of *degrees of freedom* at θ: the maximum number of linearly independent directions along which paths on the fiber can leave θ (though a path cannot necessarily use all these degrees of freedom simultaneously).

At a weight vector θ, we will construct three prebases (besides the one-matrix prebasis): one that spans $T_\theta S$, and two that span $\text{null}\, d\mu(\theta)$. (Unlike the one-matrix prebasis, they usually do not span the entire weight space $\mathbb{R}^{d_\theta}$.)

- The *freedom prebasis* spans $\text{null}\, d\mu(\theta)$. The freedom prebasis contains all the one-matrix subspaces that lie in $\text{null}\, d\mu(\theta)$ (i.e., subspaces whose moves stay on the fiber) and excludes all the one-matrix subspaces that do not (i.e., their moves leave the fiber). It also contains some two-matrix subspaces that represent directions tangent to curved paths on S, as discussed in Section 6.2.

- The *stratum prebasis* spans $T_\theta S$. The stratum prebasis contains all the one-matrix subspaces that lie in $T_\theta S$ (i.e., subspaces whose moves stay on $\bar{S}$) and excludes all the one-matrix subspaces that do not (i.e., their moves leave $\bar{S}$). Thus it omits all the one-matrix subspaces that the freedom prebasis omits, and it also omits all the combinatorial subspaces, associated with connecting and swapping moves. It contains all the two-matrix subspaces in the freedom prebasis and adds some more.

- The *fiber prebasis* is a superset of the stratum prebasis that spans $\text{null}\, d\mu(\theta)$. The fiber prebasis has fewer one-matrix subspaces and more two-matrix subspaces than the freedom prebasis, so it is a bit less elegant. But because it includes the stratum prebasis, the fiber prebasis is better "aligned" with the fiber: for every stratum S' whose closure contains θ, if $T_\theta \bar{S}'$ exists then some subset of the fiber prebasis spans $T_\theta \bar{S}'$. Thus the fiber prebasis is a single prebasis whose subspaces can be used to span the tangent spaces of the strata that meet at θ.

Although $T_\theta S \subseteq \text{null}\, d\mu(\theta)$ and the freedom prebasis spans $d\mu(\theta)$, the freedom prebasis does not, in general, have a subset whose span is $T_\theta S$. Thus the stratum prebasis must add some additional two-matrix subspaces that lie in $T_\theta S$, as well as drop the one-matrix subspaces that do not lie in $T_\theta S$. The fiber prebasis retains all the subspaces in the stratum prebasis and brings back some (but not all) of the one-matrix subspaces from the freedom prebasis, so that the fiber prebasis also spans $\text{null}\, d\mu(\theta)$, but it represents every stratum meeting at θ.

Before we can formally define the three prebases (in Section 8.4), we must introduce the differential map $d\mu$ (in Section 8.1) and the two-matrix subspaces (in Sections 8.2 and 8.3).

8.1   The Differential Map dμ 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. 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; to say it is zero is to say all of its components are 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, Trager, Kohn, and Bruna [21] use differential geometry. Given a neural network architecture $\mu : \mathbb{R}^{d_\theta} \to \mathbb{R}^{d_h \times d_0}$ and a specified weight vector $\theta \in \mathbb{R}^{d_\theta}$, the *differential map* $d\mu(\theta) : \mathbb{R}^{d_\theta} \to \mathbb{R}^{d_h \times d_0}$ is a linear map from weight space to the space of $d_h \times d_0$ matrices. We emphasize the linearity; think of the differential map as the linear term in a Taylor expansion of μ about θ. Usually we will write its argument as $\Delta\theta$, and apply the map as $\Delta W = d\mu(\theta)(\Delta\theta)$. 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 $d\mu$ denotes a map from a weight vector θ to a linear map. This might be confusing if you haven't seen it before—a map that produces a map—and it accounts for the odd notation $d\mu(\theta)(\Delta\theta)$.

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

$$d\mu(\theta)(\Delta\theta) = \sum_{j=1}^{L} W_{L\sim j} \Delta W_j W_{j-1\sim 0} = \Delta W_L W_{L-1\sim 0} + W_L \Delta W_{L-1} W_{L-2\sim 0} + \ldots + W_{L\sim 1} \Delta W_1. \qquad (8.1)$$

With $\Delta\theta$ fixed, this is the directional derivative $\mu'_{\Delta\theta}(\theta) = d\mu(\theta)(\Delta\theta)$. (But we usually like to think of θ as fixed and $\Delta\theta$ as varying. Observe that (8.1) is linear in $\Delta\theta$ but certainly not linear in θ.) As μ is continuous and smooth, if you walk from θ along a smooth path on the fiber $\mu^{-1}(W)$, your initial direction of motion $\Delta\theta$ has directional derivative zero; so your initial direction is in the nullspace of $d\mu(\theta)$, defined to be

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

Consider the set of directions by which a smooth path can leave θ on the fiber. We will show (Corollary 41 in Section 8.5) that the span of this set of directions is null $d\mu(\theta)$. Hence, we say that the number of *degrees of freedom* on the fiber at θ is the dimension of null $d\mu(\theta)$. However, not every direction in null $d\mu(\theta)$ is associated with some path leaving θ on the fiber. In Figure 3, for example, for any point θ on the stratum S_{010}, null $d\mu(\theta)$ is the entire three-dimensional weight space, and paths can leave θ on the line S_{010}, the plane S_{011}, or the plane S_{110}. However, paths cannot leave θ in all directions; they are restricted to the two planes. To help identify the directions by which a smooth path can leave θ, we will specify a *fiber prebasis* that spans null $d\mu(\theta)$ in Section 8.4.

Lemma 33. *Let $\theta \in \mathbb{R}^{d_\theta}$ be a weight vector in a stratum S in the rank stratification of a fiber $\mu^{-1}(W)$. Then every smooth path containing θ on the fiber is tangent to* null $d\mu(\theta)$ *at θ. Moreover, $T_\theta S \subseteq$* null $d\mu(\theta)$.

Proof. Let $P \subset \mu^{-1}(W)$ be a smooth path on the fiber. Every point $\zeta \in P$ satisfies $\mu(\zeta) = W$. As μ is a continuous, smooth function of $\mathbb{R}^{d_\theta}$, any vector $\Delta\zeta$ tangent to P at ζ is a direction in which μ has a directional derivative of zero—that is, $\mu'_{\Delta\zeta}(\zeta) = d\mu(\zeta)(\Delta\zeta) = 0$. Hence $\Delta\zeta \in$ null $d\mu(\zeta)$, verifying that every smooth path containing θ on the fiber is tangent to null $d\mu(\theta)$ at θ.

As S is a smooth manifold, for every nonzero vector $\Delta\theta \in T_\theta S$, there is a smooth path leaving θ on S that is tangent to $\Delta\theta$ at θ. Therefore, $T_\theta S \subseteq$ null $d\mu(\theta)$. $\qquad \square$

Recall from Section 6.1 that Θ_O^{fiber} contains all the one-matrix subspaces whose displacements stay on the fiber, and Θ_O^{L0} contains all the one-matrix subspaces whose displacements do not stay on the fiber. The displacements in Θ_O^{fiber} also stay on null $d\mu(\theta)$, as a direct corollary to the first claim of Lemma 33. We now prove it a second way, which also permits us to show that the nonzero displacements in Θ_O^{L0} are not in null $d\mu(\theta)$: by plugging the displacements into the formula (8.1).

Lemma 34. *Every subspace* $\phi_{lkjih} \in \Theta_O^{\text{fiber}}$ *satisfies* $\phi_{lkjih} \subseteq$ null d$\mu(\theta)$. *Moreover,* span $\Theta_O^{\text{fiber}} \subseteq$ null d$\mu(\theta)$. *By constrast, every subspace* $\phi_{lkjih} \in \Theta_O^{L0}$ *satisfies* $\phi_{lkjih} \cap$ null d$\mu(\theta) = \{\mathbf{0}\}$. *(That is,* ϕ_{lkjih} *and* null d$\mu(\theta)$ *are linearly independent.)*

Proof. Consider a displacement $\Delta\theta \in \phi_{lkjih} = a_{lji} \otimes b_{k,j-1,h}$. We can write $\Delta\theta = (\ldots, 0, \Delta W_j, 0, \ldots)$.

If $\phi_{lkjih} \in \Theta_O^{\text{fiber}}$, then either $L > l$ or $h > 0$ by the definition of Θ_O^{fiber}. In the former case, $W_{L\sim j}\Delta W_j = 0$ because row $\Delta W_j \subseteq a_{lji} \subseteq A_{lji} \subseteq$ null $W_{l+1\sim j} \subseteq$ null $W_{L\sim j}$. In the latter case, $\Delta W_j W_{j-1\sim0} = 0$ because col $\Delta W_j \subseteq b_{k,j-1,h} \subseteq B_{k,j-1,h} \subseteq$ null $W_{j-1\sim h-1}^\top \subseteq$ null $W_{j-1\sim0}^\top$. In both cases, by the formula (8.1), d$\mu(\theta)(\Delta\theta) = W_{L\sim j}\Delta W_j W_{j-1\sim0} = 0$. Hence $\phi_{lkjih} \subseteq$ null d$\mu(\theta)$.

It follows that span $\Theta_O^{\text{fiber}} \subseteq$ null d$\mu(\theta)$ because null d$\mu(\theta)$ is a subspace and every subspace in Θ_O^{fiber} is a subset of null d$\mu(\theta)$.

By contrast, if $\phi_{lkjih} \in \Theta_O^{L0}$, then $L = l$ and $h = 0$ by the definition of Θ_O^{L0}, so $\Delta\theta \in a_{Lji} \otimes b_{k,j-1,0}$. By Lemma 6, $W_{L\sim j}a_{Lji}$ has the same dimension as a_{Lji}; therefore, $W_{L\sim j}$ induces a linear bijection from a_{Lji} to a_{LLi}. Also by Lemma 6, $W_{j-1\sim0}^\top b_{k,j-1,0}$ has the same dimension as $b_{k,j-1,0}$, thus $W_{j-1\sim0}^\top$ induces a linear bijection from $b_{k,j-1,0}$ to b_{k00}. Hence $W_{L\sim j}\Delta W_j W_{j-1\sim0}$ is a linear bijection from $\Delta W_j \in a_{Lji} \otimes b_{k,j-1,0}$ to $a_{LLi} \otimes b_{k00}$ that maps $\Delta W_j = 0$ to zero. It follows that if $\Delta\theta \neq \mathbf{0}$, then d$\mu(\theta)(\Delta\theta) = W_{L\sim j}\Delta W_j W_{j-1\sim0} \neq 0$. Hence $\phi_{lkjih} \cap$ null d$\mu(\theta) = \{\mathbf{0}\}$ as claimed. $\qquad\square$

Compare Lemma 34 to Corollary 21. Our partition of Θ_O into Θ_O^{fiber} and Θ_O^{L0} separates the one-matrix subspaces into those in null d$\mu(\theta)$ and those linearly independent of null d$\mu(\theta)$ (Lemma 34); it also separates the one-matrix subspaces into those whose displacements stay on the fiber and those whose displacements move off the fiber (Corollary 21). This is worth remarking on, because as we have said several times, not all displacements in null d$\mu(\theta)$ are tangent to paths leaving θ on the fiber. But all displacements *in all one-matrix subspaces* in null d$\mu(\theta)$ (i.e., the subspaces in Θ_O^{fiber}) are tangent to (straight) paths leaving θ on the fiber.

Consider again Figure 3, where at the weight vector $\theta = \mathbf{0}$ (labeled S_{000}), null d$\mu(\theta)$ is the entire weight space $\mathbb{R}^3$, but smooth paths on the fiber can leave θ only along some special directions. Among those special directions are the nonzero displacements in the one-matrix subspaces ϕ_{10110}, ϕ_{21221}, and ϕ_{32332}; these displacements move from S_{000} onto the strata S_{001}, S_{010}, and S_{100}, respectively. Also among those special directions are displacements in $\phi_{10110} + \phi_{21221}$, in $\phi_{21221} + \phi_{32332}$, or in $\phi_{32332} + \phi_{10110}$, which move from S_{000} onto the strata S_{011}, S_{110}, and S_{101}, respectively.

Knowing that span $\Theta_O^{\text{fiber}} \subseteq$ null d$\mu(\theta)$, we are motivated to write a more explicit expression for span Θ_O^{fiber}. Lemma 20 states that O_j^{fiber} is a prebasis for N_j, so

$$\text{span } O_j^{\text{fiber}} = N_j = \text{null } W_{L\sim j} \otimes \mathbb{R}^{d_{j-1}} + \mathbb{R}^{d_j} \otimes \text{null } W_{j-1\sim0}^\top.$$

Therefore,

$$\text{span } \Theta_O^{\text{fiber}} = \{(\Delta W_L, \Delta W_{L-1}, \ldots, \Delta W_1) : \Delta W_j \in N_j\}. \tag{8.2}$$

As span $\Theta_O^{\text{fiber}} \subseteq$ null d$\mu(\theta)$, we will use (8.2) to derive an explicit expression for null d$\mu(\theta)$ in Section 8.5.

Recall identity (6.1) from Section 6.1, dim $N_j = d_j d_{j-1} - \text{rk } W_{L\sim j} \cdot \text{rk } W_{j-1\sim0}$. It follows that

$$\text{dim span } \Theta_O^{\text{fiber}} = \sum_{j=1}^{L}(d_j d_{j-1} - \text{rk } W_{L\sim j} \cdot \text{rk } W_{j-1\sim0}) = d_\theta - \sum_{j=1}^{L} \text{rk } W_{L\sim j} \cdot \text{rk } W_{j-1\sim0}. \tag{8.3}$$

In the two-matrix case ($L = 2$), span $\Theta_O^{\text{fiber}} = \{(\Delta W_2, \Delta W_1) : \Delta W_2 \in \mathbb{R}^{d_2} \otimes \text{null } W_1^\top, \Delta W_1 \in \text{null } W_2 \otimes \mathbb{R}^{d_0}\}$ and dim span $\Theta_O^{\text{fiber}} = d_\theta - d_2 \cdot \text{rk } W_1 - d_0 \cdot \text{rk } W_2$.

8.2　Two-Matrix Subspaces

Usually Θ_O^{fiber} does not suffice to span null $d\mu(\theta)$, and $\Theta_O^{\text{stratum}}$ does not suffice to span $T_\theta S$. To give a complete prebasis for null $d\mu(\theta)$ and a complete prebasis for $T_\theta S$, we add some two-matrix subspaces. Recall the *two-matrix paths* we defined in Section 6.2, which start at θ and lie on the stratum S that contains θ. Two-matrix paths are often curved but always smooth. The *two-matrix subspaces* we construct in this section represent some of the directions by which two-matrix paths leave θ, hence these subspaces are tangent to S and lie in $T_\theta S$. By Lemma 33, they also lie in null $d\mu(\theta)$.

A direction tangent to a two-matrix path has at most 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 (8.1) is zero. If we take a displacement $\Delta\theta$ tangent to a two-matrix path P at θ—that is, $\Delta\theta \in T_\theta P$—there are two terms in the summation (8.1) that might be nonzero, but their sum is zero.

Recall from (6.3) in Section 6.2 that the vectors tangent to P at $\theta = (W_L, W_{L-1}, \ldots, W_1)$ have the form

$$\Delta\theta = (0, 0, \ldots, 0, \Delta W_{j+1}, \Delta W_j, 0, \ldots, 0) = (0, 0, \ldots, 0, W_{j+1}H, -HW_j, 0, \ldots, 0).$$

As $P \subset S$, $\Delta\theta$ is tangent to $T_\theta S$ (assuming $\Delta\theta \neq \mathbf{0}$).

A fruitful way to define subspaces tangent to S is to consider the matrices $H \in a_{lji} \otimes b_{kjh}$, where a_{lji} and b_{kjh} are prebasis subspaces, as defined in Section 4.3. Recall that this means that col $H \subseteq a_{lji}$ and row $H \subseteq b_{kjh}$. For all l, k, j, i, and h that satisfy $L > j > 0$, $L \geq l \geq j \geq i \geq 0$, and $L \geq k \geq j \geq h \geq 0$, we define a *two-matrix subspace*

$$\tau_{lkjih} = \{(0, \ldots, 0, \underbrace{W_{j+1}H}_{\Delta W_{j+1}}, \underbrace{-HW_j}_{\Delta W_j}, 0, \ldots, 0) : H \in a_{lji} \otimes b_{kjh}\}. \tag{8.4}$$

It is important that we choose flow prebasis subspaces (rather than choosing, say, the standard prebases), as Lemma 7 says we always can, so that the following properties hold.

Lemma 35. *If we use flow prebases (associated with a point θ), each displacement $\Delta\theta \in \tau_{lkjih}$ satisfies*

$$\Delta W_{j+1} = W_{j+1}H \in \begin{cases} a_{l,j+1,i} \otimes b_{kjh} = o_{l,k,j+1,i,h}, & l > j \\ \{0\}, & l = j \end{cases} \quad and$$

$$\Delta W_j = -HW_j \in \begin{cases} a_{lji} \otimes b_{k,j-1,h} = o_{lkjih}, & j > h \\ \{0\}, & j = h. \end{cases}$$

for some $H \in a_{lji} \otimes b_{kjh}$. Moreover,

- *if $l = j = h$, then $\tau_{lkjih} = \{\mathbf{0}\}$.*
- *If $l > j = h$, then $\tau_{lkjih} = \phi_{l,k,j+1,i,h}$.*
- *If $l = j > h$, then $\tau_{lkjih} = \phi_{lkjih}$.*
- *If $l > j > h$, then no two displacements in τ_{lkjih} have the same value of ΔW_{j+1}, nor the same value of ΔW_j. Hence, each displacement in τ_{lkjih} is a sum of a unique member of $\phi_{l,k,j+1,i,h}$ and a unique member of ϕ_{lkjih}. Moreover, in every nonzero $\Delta\theta \in \tau_{lkjih}$, both ΔW_{j+1} and ΔW_j are nonzero.*

The dimension of τ_{lkjih} is $(\dim a_{lji}) \cdot (\dim b_{kjh}) = \omega_{li}\, \omega_{kh}$ in all cases except $l = j = h$, for which $\dim \tau_{lkjih} = 0$.

Proof. If $l = j$ then $W_{j+1}H = 0$ because col $H \subseteq a_{jji} \subseteq A_{jji} \subseteq$ null W_{j+1}. Whereas if $l > j$, then $W_{j+1}H \in a_{l,j+1,i} \otimes b_{kjh}$ because $H \in a_{lji} \otimes b_{kjh}$ and we are using flow prebases, implying that $W_{j+1}a_{lji} = a_{l,j+1,i}$.

Symmetrically, if $j = h$ then $-HW_j = 0$ because row $H \subseteq b_{kjj} \subseteq B_{kjj} \subseteq$ null $W_j^\top$. Whereas if $j > h$, then $-HW_j \in a_{lji} \otimes b_{k,j-1,h}$ because for flow prebases, $W_j^\top b_{kjh} = b_{k,j-1,h}$.

It follows immediately that $\tau_{lkjih} = \{\mathbf{0}\}$ if $l = j = h$ (as claimed), $\tau_{lkjih} \subseteq \phi_{l,k,j+1,i,h}$ if $l > j = h$, and $\tau_{lkjih} \subseteq \phi_{lkjih}$ if $l = j > h$.

If $l > j$, then $W_{j+1}a_{lji}$ has the same dimension as a_{lji} by Lemma 6; therefore, W_{j+1} induces a linear bijection from matrices in $a_{lji} \otimes b_{kjh}$ to matrices in $a_{l,j+1,i} \otimes b_{kjh}$, and the map used in (8.4) is a linear bijection from matrices in $a_{lji} \otimes b_{kjh}$ to displacements in τ_{lkjih}. It follows that $\tau_{lkjih} = \phi_{l,k,j+1,i,h}$ if $l > j = h$, as claimed; and if $l > j > h$, it follows that no two displacements in τ_{lkjih} have the same value of ΔW_{j+1}, as claimed.

Symmetrically, if $j > h$, then $W_j^\top b_{kjh}$ has the same dimension as b_{kjh} by Lemma 6; therefore, $W_j^\top$ induces a linear bijection from matrices in $a_{lji} \otimes b_{kjh}$ to matrices in $a_{lji} \otimes b_{k,j-1,h}$, and again the map used in (8.4) is a linear bijection from matrices in $a_{lji} \otimes b_{kjh}$ to displacements in τ_{lkjih}. Thus $\tau_{lkjih} = \phi_{lkjih}$ if $l = j > h$, as claimed; and if $l > j > h$, no two displacements in τ_{lkjih} have the same value of ΔW_j, as claimed.

If $l > j$ or $j > h$, the lienar bijection in (8.4) implies that τ_{lkjih} has the same dimension as $a_{lji} \otimes b_{kjh}$, which is $(\dim a_{lji}) \cdot (\dim b_{kjh})$.

If $l > j > h$, it follows that $\Delta W_{j+1} = 0$ only for $H = 0$, and likewise that $\Delta W_j = 0$ only for $H = 0$. This verifies that in every nonzero $\Delta\theta \in \tau_{lkjih}$, both ΔW_{j+1} and ΔW_j are nonzero. $\qquad\square$

Lemma 35 shows that some of the two-matrix subspaces are one-matrix subspaces or the trivial subspace $\{\mathbf{0}\}$. Here, we are mainly interested in the other ones—the two-matrix subspaces with $l > j > h$, the ones that typically have two nonzero matrices! Sometimes a subspace τ_{lkjih} with $l > j > h$ can be the trivial subspace $\{\mathbf{0}\}$, but only if $a_{lji} = \{\mathbf{0}\}$ or $b_{kjh} = \{\mathbf{0}\}$. We define two sets: the set Θ_{T+} of all two-matrix subspaces except $\{\mathbf{0}\}$ (typically including some one-matrix subspaces), and the set Θ_T of two-matrix subspaces for which $l > j > h$ except $\{\mathbf{0}\}$ (excluding all the one-matrix subspaces).

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

For any subspace $\tau_{lkjih} \in \Theta_T$, Lemma 35 shows that each displacement in τ_{lkjih} is a sum of two one-matrix displacements in $\phi_{l,k,j+1,i,h}$ and ϕ_{lkjih}. More simply, $\tau_{lkjih} \subseteq \phi_{lkjih} + \phi_{l,k,j+1,i,h}$. We think this is elegant. This fact motivates why we choose (8.4) to define the two-matrix subspaces: it will make it easy to prove the linear independence of the subspaces in the freedom, stratum, and fiber prebases.

For the next three lemmas, consider a weight vector $\theta = (W_L, W_{L-1}, \ldots, W_1)$ on a stratum S and the set Θ_{T+} of two-matrix subspaces associated with θ. These lemmas show that every $\tau_{lkjih} \in \Theta_{T+}$ is a subset of both $T_\theta S$ and null $d\mu(\theta)$. Hence τ_{lkjih} is tangent to S at θ.

Lemma 36. *For every two-matrix subspace $\tau_{lkjih} \in \Theta_{T+}$ and every nonzero displacement $\Delta\theta \in \tau_{lkjih}$, there exists a smooth two-matrix path $P \subset S$ that leaves θ in the direction $\Delta\theta$ (that is, $\Delta\theta \in T_\theta P$).*

Proof. By the definition (8.4) of τ_{lkjih}, $\Delta\theta = (0, 0, \ldots, 0, W_{j+1}H, -HW_j, 0, \ldots, 0)$ for some $H \in a_{lji} \otimes b_{kjh}$. Consider the two-matrix path

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

where $\hat\epsilon > 0$ is sufficiently small that $I + \epsilon H$ is invertible for all $\epsilon \in [0, \hat\epsilon]$. 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. Recall from Section 6.2 the formula (6.3) for the line tangent to P at θ, $T_\theta P = \{(0, 0, \ldots, 0, \gamma W_{j+1}H, -\gamma HW_j, 0, \ldots, 0) : \gamma \in \mathbb{R}\}$. Clearly $\Delta\theta \in T_\theta P$, so P is tangent to $\Delta\theta$ at θ. $\qquad\square$

Lemma 37. *Every $\tau_{lkjih} \in \Theta_{T+}$ satisfies $\tau_{lkjih} \subseteq T_\theta S$. Moreover, span $\Theta_T \subseteq$ span $\Theta_{T+} \subseteq T_\theta S$.*

Proof. Consider a subspace $\tau_{lkjih} \in \Theta_{T+}$ and a nonzero displacement $\Delta\theta \in \tau_{lkjih}$. By Lemma 36, there is a smooth two-matrix path $P \subset S$ with endpoint θ such that $\Delta\theta \in T_\theta P$. As S is a smooth manifold, $T_\theta P \subseteq T_\theta S$. Hence $\Delta\theta \in T_\theta S$ for every $\Delta\theta \in \tau_{lkjih}$, so $\tau_{lkjih} \subseteq T_\theta S$.

It follows that span $\Theta_{T+} \subseteq T_\theta S$ because $T_\theta S$ is a subspace and $\tau_{lkjih} \subseteq T_\theta S$ for every $\tau_{lkjih} \in \Theta_{T+}$. $\qquad\square$

An immediate corollary of Lemmas 37 and 33 is that span $\Theta_{T+} \subseteq$ null $d\mu(\theta)$. As we did with Lemma 34, we now prove it a second way: by plugging the displacements into the formula (8.1) and showing that the result is always zero.

Lemma 38. *Every $\tau_{lkjih} \in \Theta_{T+}$ satisfies $\tau_{lkjih} \subseteq$ null $d\mu(\theta)$. Moreover, span $\Theta_T \subseteq$ span $\Theta_{T+} \subseteq$ null $d\mu(\theta)$.*

Proof. Consider a subspace $\tau_{lkjih} \in \Theta_{T+}$ and a displacement $\Delta\theta \in \tau_{lkjih}$. By (8.4), there exists some $H \in a_{lji} \otimes b_{kjh}$ such that $\Delta\theta = (0, 0, \ldots, 0, W_{j+1}H, -HW_j, 0, \ldots, 0)$. By the formula (8.1), $d\mu(\theta)(\Delta\theta) = W_{L\sim j+1}\Delta W_{j+1}W_{j\sim 0} + W_{L\sim j}\Delta W_j W_{j-1\sim 0} = W_{L\sim j+1}W_{j+1}HW_{j\sim 0} - W_{L\sim j}HW_j W_{j-1\sim 0} = 0$. Hence $\Delta\theta \in$ null $d\mu(\theta)$ for all $\Delta\theta \in \tau_{lkjih}$, hence $\tau_{lkjih} \subseteq$ null $d\mu(\theta)$.

It follows that span $\Theta_{T+} \subseteq$ null $d\mu(\theta)$ because null $d\mu(\theta)$ is a subspace and $\tau_{lkjih} \subseteq$ null $d\mu(\theta)$ for every $\tau_{lkjih} \in \Theta_{T+}$. $\qquad\square$

Knowing that span $\Theta_{T+} \subseteq T_\theta S \subseteq$ null $d\mu(\theta)$ motivates us to write an explicit formula for span Θ_{T+}.

$$
\begin{aligned}
\text{span } \Theta_{T+} \;&=\; \sum_{j=1}^{L-1}\left\{(0, \ldots, 0, \underbrace{W_{j+1}H_j}_{\Delta W_{j+1}}, \underbrace{-H_j W_j}_{\Delta W_j}, 0, \ldots, 0) : H_j \in \sum_{l=j}^{L}\sum_{k=j}^{L}\sum_{i=0}^{j}\sum_{h=0}^{j} a_{lji} \otimes b_{kjh}\right\} \\
&=\; \Big\{(W_L H_{L-1}, W_{L-1}H_{L-2} - H_{L-1}W_{L-1}, \ldots, W_j H_{j-1} - H_j W_j, \ldots, \\
&\qquad\quad W_2 H_1 - H_2 W_2, -H_1 W_1) : H_j \in \mathbb{R}^{d_j \times d_j}\Big\}.
\end{aligned}
\tag{8.5}
$$

This subspace is the span of all vectors tangent to two-matrix paths at θ.

Observe that the prebasis subspaces a_{lji} and b_{kjh} disappear from (8.5), and span Θ_{T+} does not depend on our specific choices of the prebasis subspaces. By contrast, span Θ_T does depend on those choices, and we cannot write a formula for span Θ_T that is independent of them.

8.3   The Two-Matrix Subspaces We Care about Most

Consider a two-matrix subspace $\tau_{lkjih} \in \Theta_T$. Recall again that Lemma 35 shows that $\tau_{lkjih} \subseteq \phi_{lkjih} + \phi_{l,k,j+1,i,h}$. If $\phi_{lkjih} \subseteq T_\theta S$ and $\phi_{l,k,j+1,i,h} \subseteq T_\theta S$, then τ_{lkjih} adds nothing useful to the one-matrix subspaces. But if $\phi_{lkjih} \nsubseteq T_\theta S$ and $\phi_{l,k,j+1,i,h} \nsubseteq T_\theta S$ (which implies that moves with displacements in ϕ_{lkjih} or $\phi_{l,k,j+1,i,h}$ leave the stratum S), then τ_{lkjih} is interesting and useful, because $\tau_{lkjih} \subseteq T_\theta S$ (by Lemma 37). If, moreover, $\phi_{lkjih} \nsubseteq$ null $d\mu(\theta)$ and $\phi_{l,k,j+1,i,h} \nsubseteq$ null $d\mu(\theta)$ (which implies that the corresponding moves leave the fiber), τ_{lkjih} is even more interesting because $\tau_{lkjih} \subseteq$ null $d\mu(\theta)$ (by Lemma 38). Think of the latter case as the circumstance where the expression (8.1) is zero because every two-matrix displacement $\Delta\theta \in \tau_{lkjih}$ satisfies $W_{L\sim j+1}\Delta W_{j+1}W_{j\sim 0} + W_{L\sim j}\Delta W_j W_{j-1\sim 0} = 0$, but those two terms are nonzero.

As every two-matrix subspace satisfies $\tau_{lkjih} \subseteq T_\theta S \subseteq$ null $d\mu(\theta)$, we will use τ_{lkjih} in the freedom, stratum, and fiber prebases if ϕ_{lkjih} and $\phi_{l,k,j+1,i,h}$ do not lie in null $d\mu(\theta)$; and we will use τ_{lkjih} in the stratum and fiber

prebases if ϕ_{lkjih} and $\phi_{l,k,j+1,i,h}$ do not lie in $T_\theta S$. Recall from Corollary 21 in Section 7.2 that one-matrix moves with nonzero displacements in ϕ_{lkjih} or $\phi_{l,k,j+1,i,h}$ leave the fiber if and only if $l = L$ and $h = 0$. Recall from Section 7.4 that one-matrix moves with displacements in ϕ_{lkjih} or $\phi_{l,k,j+1,i,h}$ 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.) Hence we define the sets

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

For example, in the two-matrix case ($L = 2$),

$$
\begin{aligned}
\Theta_{\mathrm{T}}^{L0} = \Theta_{\mathrm{T}} &= \{\tau_{21100}, \tau_{21110}, \tau_{22100}, \tau_{22110}\} \setminus \{\{\mathbf{0}\}\} \quad \text{and} \\
\Theta_{\mathrm{T}}^{\mathrm{comb}} &= \{\tau_{21110}\} \setminus \{\{\mathbf{0}\}\}.
\end{aligned}
$$

The freedom, stratum, and fiber prebases will include Θ_{T}^{L0} (as a subset). The stratum and fiber prebases will also include $\Theta_{\mathrm{T}}^{\mathrm{comb}}$. We use the notation $\Theta_{\mathrm{T}}^{\mathrm{comb}}$ because $\Theta_{\mathrm{T}}^{\mathrm{comb}}$ replaces $\Theta_{\mathrm{O}}^{\mathrm{comb}}$ in the stratum prebasis, but it is a bit of a misnomer, as the subspaces in $\Theta_{\mathrm{T}}^{\mathrm{comb}}$ do not represent combinatorial moves.

As $\tau_{lkjih} \subseteq \phi_{lkjih} + \phi_{l,k,j+1,i,h}$ for each $\tau_{lkjih} \in \Theta_{\mathrm{T}}$, we have span $\Theta_{\mathrm{T}}^{L0} \subseteq$ span Θ_{O}^{L0} and span $\Theta_{\mathrm{T}}^{\mathrm{comb}} \subseteq$ span $\Theta_{\mathrm{O}}^{\mathrm{comb}}$. However, by Lemma 37,

$$
\text{span} \left(\Theta_{\mathrm{T}}^{L0} \cup \Theta_{\mathrm{T}}^{\mathrm{comb}}\right) \subseteq T_\theta S \subseteq \text{null } \mathrm{d}\mu(\theta),
$$

whereas no subspace in Θ_{O}^{L0} is a subset of null $\mathrm{d}\mu(\theta)$ and no subspace in Θ_{O}^{L0} nor $\Theta_{\mathrm{O}}^{\mathrm{comb}}$ is a subset of $T_\theta S$.

To understand a little more deeply the relationship between $\Theta_{\mathrm{T}}^{\mathrm{comb}}$ and $\Theta_{\mathrm{O}}^{\mathrm{comb}}$, consider a subspace $\tau_{lkjih} \in \Theta_{\mathrm{T}}^{\mathrm{comb}}$. A nonzero displacement $\Delta\theta \in \tau_{lkjih}$ is a sum of a displacement in ϕ_{lkjih} and a displacement in $\phi_{l,k,j+1,i,h}$, both of which are subspaces in $\Theta_{\mathrm{O}}^{\mathrm{comb}}$. By the definition of $\Theta_{\mathrm{T}}^{\mathrm{comb}}$, τ_{lkjih} has $l > k \geq i > h$, so the displacement in ϕ_{lkjih} corresponds to a swapping move (never a connecting move) that moves from the stratum S that contains θ to another stratum S', and the displacement in $\phi_{l,k,j+1,i,h}$ corresponds to a *different* swapping move from S to the *same* stratum S'—so it is the same *kind* of swapping move, but it changes ΔW_{j+1} instead of ΔW_j. (Specifically, S' is 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 S). The displacement in ϕ_{lkjih} and the displacement in $\phi_{l,k,j+1,i,h}$ are balanced so that $\Delta\theta$ is tangent to S. We can think of $\Theta_{\mathrm{T}}^{\mathrm{comb}}$ as a set of degrees of freedom along S that cannot be expressed as one-matrix subspaces because changing only one matrix would trigger a swapping move. Similarly, we can think of Θ_{T}^{L0} as a set of degrees of freedom along S that cannot be expressed as one-matrix subspaces because changing only one matrix would change the value of W.

8.4   The Freedom, Stratum, and Fiber Prebases

The *freedom prebasis* at θ is a set of linearly independent subspaces of $\mathbb{R}^{d_\theta}$ that spans null $\mathrm{d}\mu(\theta)$, namely,

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

The freedom prebasis contains every one-matrix subspace $\phi_{lkjih} \neq \{\mathbf{0}\}$ such that $\phi_{lkjih} \subseteq$ null $\mathrm{d}\mu(\theta)$, plus some additional two-matrix subspaces as needed so that Θ^{free} spans null $\mathrm{d}\mu(\theta)$ (as we will show). The freedom

prebasis excludes the subspaces ϕ_{Lkji0} because they do not lie in null $d\mu(\theta)$, but it contains the two-matrix subspaces τ_{Lkji0}, which lie in null $d\mu(\theta)$ by Lemma 38.

Recall from Section 8.3 that span $\Theta_T^{L0} \subseteq$ span Θ_O^{L0}. But no subspace in Θ_O^{L0} is a subset of null $d\mu(\theta)$, whereas span $\Theta_T^{L0} \subseteq$ null $d\mu(\theta)$. The definition of Θ^{free} above takes the set Θ_O of one-matrix subspaces, removes the subspaces in Θ_O^{L0}, and replaces them with the two-matrix subspaces in Θ_T^{L0}. Specifically, for each pair $k, i \in [0, L]$ satisfying $k + 1 \geq i$, the definition removes the subspaces ϕ_{Lkji0} for $j \in [i, k+1]$, and replaces them with the subspaces τ_{Lkji0} for $j \in [i, k]$. Thus it removes $k - i + 2$ one-matrix subspaces of dimension $\omega_{Li} \omega_{k0}$ and replaces them with $k - i + 1$ two-matrix subspaces, also of dimension $\omega_{Li} \omega_{k0}$.

Let $W = \mu(\theta)$ and let S be the stratum of $\mu^{-1}(W)$ that contains θ. The *stratum prebasis* at θ is a set of linearly independent subspaces of $\mathbb{R}^{d_\theta}$ that spans the tangent space $T_\theta S$, namely,

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

The stratum prebasis contains every one-matrix subspace $\phi_{lkjih} \neq \{\mathbf{0}\}$ such that $\phi_{lkjih} \subseteq T_\theta S$, plus some additional two-matrix subspaces as needed so that Θ^{stratum} spans $T_\theta S$. The stratum prebasis, like the freedom prebasis, excludes the one-matrix subspaces ϕ_{Lkji0} (because their moves move off the fiber), but it also excludes all the combinatorial one-matrix subspaces (because their moves move off the stratum S). The stratum prebasis, like the freedom prebasis, includes the two-matrix subspaces in Θ_T^{L0}, but it also includes the two-matrix subspaces in Θ_T^{comb}.

We think of Θ^{stratum} as being obtained from Θ^{free} as follows: for each choice of indices l, k, i, h satisfying $L \geq l \geq k + 1 \geq i > h \geq 0$, the definition removes the subspaces ϕ_{lkjih} for $j \in [i, k+1]$, and replaces them with the subspaces τ_{lkjih} for $j \in [i, k]$. Thus it removes $k - i + 2$ combinatorial one-matrix subspaces of dimension $\omega_{li} \omega_{kh}$ and replaces them with $k - i + 1$ two-matrix subspaces, also of dimension $\omega_{li} \omega_{kh}$.

The *fiber prebasis* at θ is a superset of the stratum prebasis that spans null $d\mu(\theta)$, namely,

$$
\begin{aligned}
\Theta^{\text{fiber}} \quad &= \quad \Theta^{\text{stratum}} \cup \{ \phi_{l,k,k+1,i,h} \in \Theta_O : (L > l \text{ or } h > 0) \text{ and } l > k \text{ and } i > h \} \\
&= \quad \left(\Theta_O \setminus \{ \phi_{lkjih} \in \Theta_O : (l = L \text{ and } h = 0) \text{ or } (l > k \geq j \geq i > h) \} \right) \cup \\
& \qquad \{ \tau_{lkjih} \in \Theta_T : (l = L \text{ and } h = 0) \text{ or } (l > k \geq j \geq i > h) \}.
\end{aligned}
$$

Recall our goal for the fiber prebasis: for every stratum S' whose closure contains θ, if $T_\theta \bar{S}'$ exists, we want some subset of Θ^{fiber} to be a prebasis for $T_\theta \bar{S}'$. As $T_\theta \bar{S}' \supseteq T_\theta S$, that prebasis is a superset of Θ^{stratum}. The fiber prebasis gives us a compact way to express the tangent spaces of all the strata meeting at a point θ.

We think of Θ^{fiber} as being obtained from Θ^{free} as follows: for each choice of indices l, k, i, h satisfying $L \geq l \geq k + 1 \geq i > h \geq 0$, and for each $j \in [i, k]$, the definition removes ϕ_{lkjih} and replaces it with τ_{lkjih}. (Recall that ϕ_{lkjih} and τ_{lkjih} have the same dimension, consistent with our claim that span $\Theta^{\text{fiber}} =$ span Θ^{free}.) When we obtained Θ^{stratum} from Θ^{free} we also removed the combinatorial subspaces ϕ_{lkjih} for $j = k + 1$, but Θ^{free} retains them. Note that this is an arbitrary choice—for each index set l, k, i, h, we could have chosen any single $j \in [i, k+1]$ for retention—but choosing $j = k + 1$ makes the indexes nice.

In the two-matrix case ($L = 2$),

$$
\begin{aligned}
\Theta^{\text{free}} = \Theta^{\text{fiber}} \quad &= \quad \{ \phi_{10100}, \phi_{10110}, \phi_{11100}, \phi_{11110}, \phi_{12100}, \phi_{12110}, \\
& \qquad \phi_{21201}, \phi_{21211}, \phi_{21221}, \phi_{22201}, \phi_{22211}, \phi_{22221}, \tau_{21100}, \tau_{21110}, \tau_{22100}, \tau_{22110} \} \setminus \{\{\mathbf{0}\}\}, \\
\Theta^{\text{stratum}} \quad &= \quad \{ \phi_{10100}, \phi_{11100}, \phi_{11110}, \phi_{12100}, \phi_{12110}, \\
& \qquad \phi_{21201}, \phi_{21211}, \phi_{22201}, \phi_{22211}, \phi_{22221}, \tau_{21100}, \tau_{21110}, \tau_{22100}, \tau_{22110} \} \setminus \{\{\mathbf{0}\}\}.
\end{aligned}
$$

8.5 How to Prove that the Freedom, Stratum, and Fiber Prebases are Prebases

This section outlines our strategy for proving that Θ^{stratum} is a prebasis for $T_\theta S$, and that Θ^{free} and Θ^{fiber} are prebases for null $d\mu(\theta)$. Unfortunately, the proofs will not be truly complete until the end of Section 9, as they involve some counting of dimensions that we defer to Sections 8.7 and 9.4. But the incomplete proofs we give here motivate the counting.

There are three steps to each proof. The first step is to observe, as we already have in Lemmas 25, 34, 37, and 38, that every subspace in Θ^{stratum} is a subset of $T_\theta S$, and every subspace in Θ^{free} or Θ^{fiber} is a subset of null $d\mu(\theta)$. Hence span $\Theta^{\text{stratum}} \subseteq T_\theta S$, span $\Theta^{\text{free}} \subseteq$ null $d\mu(\theta)$, and span $\Theta^{\text{fiber}} \subseteq$ null $d\mu(\theta)$. The second step is to show that the subspaces in Θ^{free} are linearly independent, and so are the subspaces in Θ^{stratum} or Θ^{fiber}. This step is Lemma 39, immediately below. The third and final step is to add up the dimensions of the subspaces in Θ^{free} (or Θ^{fiber}) and see that their total dimension is the dimension of null $d\mu$; hence span $\Theta^{\text{free}} =$ null $d\mu(\theta)$ and span $\Theta^{\text{fiber}} =$ null $d\mu(\theta)$. The third step is more complicated for Θ^{stratum}, as we haven't yet derived the dimension of $T_\theta S$. The subspace orthogonal to S at θ, denoted $N_\theta S$, is the orthogonal complement of $T_\theta S$ in $\mathbb{R}^{d_\theta}$, so dim $N_\theta S +$ dim $T_\theta S = d_\theta$. In Section 9 we will identify some subspaces in $N_\theta S$, then we will find that the sum of the dimensions of those subspaces plus the sum of the dimensions of the subspaces in Θ^{stratum} is d_θ. Therefore, span $\Theta^{\text{stratum}} = T_\theta S$.

Lemma 39. *The subspaces in Θ^{free} are linearly independent. Likewise for Θ^{stratum} and Θ^{fiber}.*

Proof. Let Θ be one of Θ^{free}, Θ^{stratum}, or Θ^{fiber}. Suppose for the sake of contradiction that Θ is not linearly independent. Then we can choose one weight vector from each subspace in Θ—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 19, the one-matrix subspaces in Θ_O are linearly independent. Therefore, at least one canceling vector from a two-matrix subspace is nonzero. Let $\tau_{lkjih} \in \Theta$ be a two-matrix subspace whose canceling vector $\zeta \in \tau_{lkjih}$ is nonzero such that the index j is minimal—there is no nonzero canceling vector from a two-matrix subspace with a smaller j. The fact that $\tau_{lkjih} \in \Theta$ implies that $\phi_{lkjih} \notin \Theta$ (by the definitions of Θ^{free}, Θ^{stratum}, or Θ^{fiber}).

Given a weight vector $\xi = (X_L, X_{L-1}, \ldots, X_1)$, let $M_j(\xi) = X_j$. Given a subspace σ of weight vectors, let $M_j(\sigma) = \{M_j(\xi) : \xi \in \sigma\}$. It is clear from the definition of the one-matrix subspaces that $M_j(\phi_{vujts}) = o_{vujts}$ and $M_j(\phi_{vuj'ts}) = 0$ for $j' \neq j$. By Lemma 35, $M_j(\tau_{vujts}) = o_{vujts}$, $M_j(\tau_{v,u,j-1,t,s}) = o_{vujts}$, and $M_j(\tau_{vuj'ts}) = 0$ for $j' \notin \{j-1, j\}$. Every subspace in Θ falls into one of these five cases.

The only subspaces $\sigma \in \Theta$ satisfying $M_j(\sigma) = o_{lkjih}$ are τ_{lkjih} and $\tau_{l,k,j-1,i,h}$ (recall that $\phi_{lkjih} \notin \Theta$), and the latter might not be in Θ. If $\tau_{l,k,j-1,i,h} \in \Theta$, its cancelling vector is zero (otherwise we would have made a different choice of τ_{lkjih}). As $\zeta \neq \mathbf{0}$, by Lemma 35, $M_j(\zeta) \in o_{lkjih} \setminus \{0\}$. Every other canceling vector ξ satisfies $M_j(\xi) = 0$ or $M_j(\xi) \in o_{vujts}$ for some subspace $o_{vujts} \neq o_{lkjih}$. The one-matrix subspaces are linearly independent by Lemma 18, so the sum $\sum_\xi M_j(\xi)$ is nonzero, where ξ varies over all the canceling vectors. But by assumption the canceling vectors sum to zero, so $\sum_\xi M_j(\xi) = 0$, a contradiction. It follows that the subspaces in Θ are linearly independent. $\square$

Theorem 40. Θ^{free} *is a prebasis for* null $d\mu(\theta)$. *So is* Θ^{fiber}. *In particular,* null $d\mu(\theta) =$ span $\Theta^{\text{free}} =$ span Θ^{fiber}. *Moreover, the dimension of* null $d\mu(\theta)$ *is*

$$D^{\text{free}} = d_\theta - \sum_{j=1}^{L} \text{rk } W_{L\sim j} \cdot \text{rk } W_{j-1\sim 0} + \sum_{j=1}^{L-1} \text{rk } W_{L\sim j} \cdot \text{rk } W_{j\sim 0}. \tag{8.6}$$

Proof. By Lemmas 34 and 38, each subspace in Θ^{free} or Θ^{fiber} is a subspace of null $d\mu(\theta)$, so span $\Theta^{\text{free}} \subseteq$ null $d\mu(\theta)$ and span $\Theta^{\text{fiber}} \subseteq$ null $d\mu(\theta)$. Our goal is to show that all three subspaces have the same dimension, hence null $d\mu(\theta) = \text{span } \Theta^{\text{free}} = \text{span } \Theta^{\text{fiber}}$.

By Lemma 39, the subspaces in Θ^{free} are linearly independent. Therefore, the dimension of span Θ^{free} is the sum of the dimensions of all the subspaces in Θ^{free}. We define D^{free} to be that sum. In Section 8.7, we will show that that sum is the expression (8.6).

Trager, Kohn, and Bruna [21] show that the same expression is the dimension of null $d\mu(\theta)$. (More precisely, Trager et al. show that the image of $d\mu(\theta)$ has dimension $d_\theta - D^{\text{free}}$. It follows from the Rank-Nullity Theorem that $\dim \text{null } d\mu(\theta) = D^{\text{free}}$. See Appendix C for details and a proof.) This confirms that null $d\mu(\theta) = \text{span } \Theta^{\text{free}}$. As Θ^{free} is linearly independent, Θ^{free} is a prebasis for null $d\mu(\theta)$.

An identical argument shows that Θ^{fiber} too is a prebasis for null $d\mu(\theta)$. □

Corollary 41. *The set of directions by which a smooth path can leave θ on the fiber $\mu^{-1}(W)$ spans* null $d\mu(\theta)$. *Informally, D^{free} is the number of degrees of freedom along which a smooth path can leave θ on the fiber.*

Proof. Every point ζ on the fiber satisfies $\mu(\zeta) = W$. If there is a smooth path on the fiber leaving θ tangent to some direction $\Delta\theta$, then the directional derivative $d\mu(\theta)(\Delta\theta)$ is zero; hence $\Delta\theta \in$ null $d\mu(\theta)$.

By Corollary 21, for every nonzero displacement $\Delta\theta$ in every one-matrix subspace $\phi_{lkjih} \in \Theta^{\text{free}}$, a ray originating at θ in the direction $\Delta\theta$ is a smooth path on the fiber. By Lemma 36, for every displacement $\Delta\theta$ in every two-matrix subspace $\tau_{lkjih} \in \Theta^{\text{free}}$, there is a smooth two-matrix path on the fiber that leaves θ in the direction $\Delta\theta$. By Theorem 40, null $d\mu(\theta) = \text{span } \Theta^{\text{free}}$; hence these directions suffice to span null $d\mu(\theta)$. □

Theorem 42. Θ^{stratum} *is a prebasis for $T_\theta S$. In particular, $T_\theta S = \text{span } \Theta^{\text{stratum}}$. Moreover, the dimension of $T_\theta S$ and the dimension of S is*

$$D^{\text{stratum}} = d_\theta - \text{rk } W \cdot (d_L + d_0 - \text{rk } W) - \sum_{L \geq k+1 \geq i > 0} \beta_{k+1,i,i} \, \alpha_{k,k,i-1}. \tag{8.7}$$

Proof. By Lemma 39, the subspaces in Θ^{stratum} are linearly independent. Therefore, the dimension of span Θ^{stratum} is the sum of the dimensions of all the subspaces in Θ^{stratum}. We define D^{stratum} to be that sum. In Section 8.7, we will show that this sum is the expression (8.7).

By Lemmas 25 and 37, each subspace in Θ^{stratum} is a subspace of $T_\theta S$, so span $\Theta^{\text{stratum}} \subseteq T_\theta S$. Hence the dimension of $T_\theta S$ is at least D^{stratum}. Let $N_\theta S$ denote the orthogonal complement of $T_\theta S$ in the space $\mathbb{R}^{d_\theta}$; $N_\theta S$ is the (highest-dimensional) subspace normal to S at θ. We will show in Lemma 51 that the dimension of $N_\theta S$ is at least

$$\text{rk } W \cdot (d_L + d_0 - \text{rk } W) + \sum_{L \geq k+1 \geq i > 0} \beta_{k+1,i,i} \, \alpha_{k,k,i-1}.$$

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 $T_\theta S = \text{span } \Theta^{\text{stratum}}$. Therefore, Θ^{stratum} is a prebasis for $T_\theta S$. □

8.6 Expressions for the Tangent Space of S and the Nullspace of d$\mu(\theta)$

We can now write explicit expressions for $T_\theta S$ and null d$\mu(\theta)$. By Lemma 37, span $\Theta_{T+} \subseteq T_\theta S$, and by Lemma 38, span $\Theta_{T+} \subseteq$ null d$\mu(\theta)$, so from (7.2), (8.2), and (8.5) we have

$$
\begin{aligned}
\text{null d}\mu(\theta) \;=\;& \text{span}\,\Theta^{\text{free}} = \text{span}\,(\Theta^{\text{free}} \cup \Theta_{T+}) = \text{span}\,(\Theta_O^{\text{fiber}} \cup \Theta_{T+}) \\
=\;& \Big\{ (\Delta W_L + W_L H_{L-1}, \Delta W_{L-1} + W_{L-1} H_{L-2} - H_{L-1} W_{L-1}, \dots, \\
& \Delta W_j + W_j H_{j-1} - H_j W_j, \dots, \Delta W_2 + W_2 H_1 - H_2 W_2, \Delta W_1 - H_1 W_1) : \\
& H_j \in \mathbb{R}^{d_j \times d_j}, \Delta W_j \in \text{null } W_{L\sim j} \otimes \mathbb{R}^{d_{j-1}} + \mathbb{R}^{d_j} \otimes \text{null } W_{j-1\sim 0}^\top \Big\}, \qquad (8.8) \\
T_\theta S \;=\;& \text{span}\,\Theta^{\text{stratum}} = \text{span}\,(\Theta^{\text{stratum}} \cup \Theta_{T+}) = \text{span}\,(\Theta_O^{\text{stratum}} \cup \Theta_{T+}) \\
=\;& \Big\{ (\Delta W_L + W_L H_{L-1}, \Delta W_{L-1} + W_{L-1} H_{L-2} - H_{L-1} W_{L-1}, \dots, \\
& \Delta W_j + W_j H_{j-1} - H_j W_j, \dots, \Delta W_2 + W_2 H_1 - H_2 W_2, \Delta W_1 - H_1 W_1) : \\
& H_j \in \mathbb{R}^{d_j \times d_j}, \\
& \Delta W_j \in \sum_{h=1}^{j-1} \text{col } W_{j\sim h} \otimes \text{null } W_{j-1\sim h-1}^\top + (\text{null } W_{L\sim j} \cap \text{col } W_{j\sim 0}) \otimes \mathbb{R}^{d_{j-1}} + \\
& \sum_{l=j}^{L-1} \text{null } W_{l+1\sim j} \otimes \text{row } W_{l\sim j-1} + \mathbb{R}^{d_j} \otimes (\text{row } W_{L\sim j-1} \cap \text{null } W_{j-1\sim 0}^\top) \Big\}. \qquad (8.9)
\end{aligned}
$$

(Recall the conventions that null $W_{L\sim L} = \{\mathbf{0}\}$ and null $W_{0\sim 0}^\top = \{\mathbf{0}\}$.)

For example, in the two-matrix case ($L = 2$),

$$
\begin{aligned}
\text{null d}\mu(\theta) \;=\;& \{ (\Delta W_2 + W_2 H, \Delta W_1 - H W_1) : \\
& H \in \mathbb{R}^{d_1 \times d_1}, \Delta W_2 \in \mathbb{R}^{d_2} \otimes \text{null } W_1^\top, \Delta W_1 \in \text{null } W_2 \otimes \mathbb{R}^{d_0} \}, \\
T_\theta S \;=\;& \Big\{ (\Delta W_2 + W_2 H, \Delta W_1 - H W_1) : \\
& H \in \mathbb{R}^{d_1 \times d_1}, \Delta W_2 \in \text{col } W_2 \otimes \text{null } W_1^\top + \mathbb{R}^{d_2} \otimes (\text{row } W_2 \cap \text{null } W_1^\top), \\
& \Delta W_1 \in (\text{null } W_2 \cap \text{col } W_1) \otimes \mathbb{R}^{d_0} + \text{null } W_2 \otimes \text{row } W_1 \Big\}.
\end{aligned}
$$

We note in passing that in (8.8), we can replace "$H_j \in \mathbb{R}^{d_j \times d_j}$" with "$H_j \in$ row $W_{L\sim j} \otimes$ col $W_{j\sim 0}$." It is possible to narrow the range of H_j in (8.9) as well, though not as narrow; for instance, to "$H_j \in$ row $W_{j+1} \otimes$ col W_j," or even narrower.[6]

8.7 Counting Dimensions

Let D_O, D_O^{L0}, D_O^{fiber}, D_O^{comb}, D_O^{stratum}, D_T^{L0}, D_T^{comb}, D^{free}, D^{stratum}, and D^{fiber} (and so forth) denote the dimension of the subspace (of $\mathbb{R}^{d_\theta}$) spanned by the prebasis Θ_O, Θ_O^{L0}, Θ_O^{fiber}, Θ_O^{comb}, $\Theta_O^{\text{stratum}}$, Θ_T^{L0}, Θ_T^{comb}, Θ^{free}, Θ^{stratum}, and Θ^{fiber}, respectively, at a specified weight vector θ. Table 6 gives the definitions of these prebases and several others, and the dimensions of the subspaces they span. In this section we derive those dimensions. (See Appendix D for some additional prebases.)

---

[6]As null d$\mu(\theta)$ = span $(\Theta_O^{\text{fiber}} \cup \Theta_T^{L0})$, in (8.8) it suffices to use $H_j \in \sum_{k=j}^L \sum_{i=0}^j a_{Lji} \otimes b_{kj0}$, because this choice generates span Θ_T^{L0} in (8.8). If we use the standard prebases for a_{Lji} and b_{kj0} (flow prebases are not needed in this context), we find that $H_j \in$ row $W_{L\sim j} \otimes$ col $W_{j\sim 0}$ suffices. (Other prebases will give other valid options.) As $T_\theta S$ = span $(\Theta_O^{\text{stratum}} \cup \Theta_T^{L0} \cup \Theta_T^{\text{comb}})$, in (8.9) it suffices to use $H_j \in \sum_{k=j}^L \sum_{i=0}^j a_{Lji} \otimes b_{kj0} + \sum_{L \ge l > k \ge j} \sum_{j \ge i > h \ge 0} a_{lji} \otimes b_{kjh}$, because this choice generates span $(\Theta_T^{L0} \cup \Theta_T^{\text{comb}})$ in (8.9). Again we could use the standard prebases to write out an explicit range, albeit a messy one. A simple superset of that range is row $W_{j+1} \otimes$ col W_j.

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

$$D_{\mathrm{O}} = d_\theta = \sum_{j=1}^{L} d_j d_{j-1} \tag{8.10}$$

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

$$D_{\mathrm{O}}^{L0} = \sum_{j=1}^{L} \left(\sum_{i=0}^{j} \omega_{Li} \right) \left(\sum_{k=j-1}^{L} \omega_{k0} \right) = \sum_{j=1}^{L} \mathrm{rk}\, W_{L\sim j} \cdot \mathrm{rk}\, W_{j-1\sim 0} \tag{8.11}$$

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

$$D_{\mathrm{O}}^{\mathrm{fiber}} = D_{\mathrm{O}} - D_{\mathrm{O}}^{L0} = d_\theta - \sum_{j=1}^{L} \mathrm{rk}\, W_{L\sim j} \cdot \mathrm{rk}\, W_{j-1\sim 0} \tag{8.12}$$

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

$$D_{\mathrm{O}}^{\mathrm{comb}} = \sum_{L \geq k+1 \geq i > 0} (k-i+2) \underbrace{(\mathrm{rk}\, W_{k+1\sim i} - \mathrm{rk}\, W_{k+1\sim i-1})}_{\beta_{k+1,i,i}} \underbrace{(\mathrm{rk}\, W_{k\sim i-1} - \mathrm{rk}\, W_{k+1\sim i-1})}_{\alpha_{k,k,i-1}} \tag{8.13}$$

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

$$D_{\mathrm{O}}^{L0,\neg\mathrm{comb}} = \mathrm{rk}\, W \cdot \sum_{j=1}^{L} \left(\mathrm{rk}\, W_{L\sim j} + \mathrm{rk}\, W_{j-1\sim 0} - \mathrm{rk}\, W \right) \tag{8.14}$$

$$\Theta_{\mathrm{O}}^{\mathrm{stratum}} = \Theta_{\mathrm{O}}^{\mathrm{fiber}} \setminus \Theta_{\mathrm{O}}^{\mathrm{comb}} = \Theta_{\mathrm{O}} \setminus \Theta_{\mathrm{O}}^{L0} \setminus \Theta_{\mathrm{O}}^{\mathrm{comb}} = \{\phi_{lkjih} \in \Theta_{\mathrm{O}} : (L > l \text{ or } h > 0) \text{ and } (l \leq k \text{ or } i \leq h)\}$$

$$D_{\mathrm{O}}^{\mathrm{stratum}} = D_{\mathrm{O}} - D_{\mathrm{O}}^{\mathrm{comb}} - D_{\mathrm{O}}^{L0,\neg\mathrm{comb}} \tag{8.15}$$

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

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

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

$$D_{\mathrm{T}}^{L0} = \sum_{j=1}^{L-1} \left(\sum_{i=0}^{j} \omega_{Li} \right) \left(\sum_{k=j}^{L} \omega_{k0} \right) = \sum_{j=1}^{L-1} \mathrm{rk}\, W_{L\sim j} \cdot \mathrm{rk}\, W_{j\sim 0} \tag{8.16}$$

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

$$D_{\mathrm{T}}^{\mathrm{comb}} = \sum_{L > k \geq i > 0} (k-i+1) \underbrace{(\mathrm{rk}\, W_{k+1\sim i} - \mathrm{rk}\, W_{k+1\sim i-1})}_{\beta_{k+1,i,i}} \underbrace{(\mathrm{rk}\, W_{k\sim i-1} - \mathrm{rk}\, W_{k+1\sim i-1})}_{\alpha_{k,k,i-1}} \tag{8.17}$$

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

$$D_{\mathrm{T}}^{L0,\neg\mathrm{comb}} = \mathrm{rk}\, W \cdot \sum_{j=1}^{L-1} \left(\mathrm{rk}\, W_{L\sim j} + \mathrm{rk}\, W_{j\sim 0} - \mathrm{rk}\, W \right) \tag{8.18}$$

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

$$D^{\mathrm{free}} = D^{\mathrm{fiber}} = d_\theta - \sum_{k,i \in [0,L], k+1 \geq i} \omega_{Li}\, \omega_{k0} = d_\theta - \sum_{j=1}^{L} \mathrm{rk}\, W_{L\sim j} \cdot \mathrm{rk}\, W_{j-1\sim 0} + \sum_{j=1}^{L-1} \mathrm{rk}\, W_{L\sim j} \cdot \mathrm{rk}\, W_{j\sim 0} \tag{8.19}$$

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

$$D^{\mathrm{stratum}} = d_\theta - \mathrm{rk}\, W \cdot (d_L + d_0 - \mathrm{rk}\, W) - \sum_{L \geq k+1 \geq i > 0} \underbrace{(\mathrm{rk}\, W_{k+1\sim i} - \mathrm{rk}\, W_{k+1\sim i-1})}_{\beta_{k+1,i,i}} \underbrace{(\mathrm{rk}\, W_{k\sim i-1} - \mathrm{rk}\, W_{k+1\sim i-1})}_{\alpha_{k,k,i-1}} \tag{8.20}$$

$$\Theta^{\mathrm{fiber}} = \Theta^{\mathrm{stratum}} \cup \{\phi_{l,k,k+1,i,h} \in \Theta_{\mathrm{O}} : (L > l \text{ or } h > 0) \text{ and } l > k \text{ and } i > h\}$$

Table 6: Sets of subspaces of $\mathbb{R}^{d_\theta}$ and their total dimensions. See also Table 7.

The most important of these numbers is D^{stratum}, the dimension of span $\Theta^{\text{stratum}} = T_\theta S$ and therefore the dimension of the stratum S that contains θ; and D^{free}, the dimension of span $\Theta^{\text{free}} = \text{null} \, d\mu(\theta)$, which we interpret as the number of degrees of freedom of motion along the paths leaving θ on the fiber. We have already used these dimensions retroactively to prove Theorems 40 (showing that span $\Theta^{\text{free}} = \text{null} \, d\mu(\theta)$) and 42 (showing that span $\Theta^{\text{stratum}} = T_\theta S$) in Section 8.5. The other counts have interpretations too: for instance, D_O^{fiber} counts the degrees of freedom of one-matrix moves (straight paths) that stay on the fiber.

Admittedly, this section makes for mind-numbing reading and can be safely skipped. It serves as a reference for anyone who wants to know the dimensions of specific subspaces, to check our proofs carefully, or to extend the results and ideas in this paper.

Recall that a one-matrix subspace $\phi_{lkjih} \in \Theta_O$ or a two-matrix subspace $\tau_{lkjih} \in \Theta_T$ has dimension $\omega_{li} \, \omega_{kh}$. If the subspaces in a set Θ 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. That is, the dimension is

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

We can apply this formula to the one-matrix prebasis Θ_O or any subset of Θ_O, because its members, the one-matrix subspaces, are linearly independent by Lemma 19. As Θ_O spans the entire weight space $\mathbb{R}^{d_\theta}$ (Lemma 19 again), $D_O = \dim \text{span} \, \Theta_O = d_\theta$. By Lemma 39, each of Θ^{free}, Θ^{stratum}, Θ^{fiber} is linearly independent, so we can also apply the formula to any subset of one of them.

In Section 7.2 we defined $\Theta_O^{L0} \subseteq \Theta_O$, representing the one-matrix moves that move off the fiber (change $\mu(\theta)$), as a prelude to defining $\Theta_O^{\text{fiber}} = \Theta_O \setminus \Theta_O^{L0}$, 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 Θ_O^{fiber} contains all the other one-matrix subspaces. In Table 6, two formulae for $D_O^{L0} = \dim \text{span} \, \Theta_O^{L0}$ appear in (8.11). The first formula follows immediately from the second definition for Θ_O^{L0} above it. The second formula follows from the identity (4.9).

In Section 8.2 we defined $\Theta_T^{L0} \subseteq \Theta_T$, which contains all the two-matrix subspaces τ_{lkjih} such that $l = L$ and $h = 0$. As $\Theta_T^{L0} \subseteq \Theta^{\text{free}}$, the subspaces in Θ_T^{L0} are linearly independent (by Lemma 39), so the dimension $D_T^{L0} = \dim \text{span} \, \Theta_T^{L0}$ is equal to the sum of the dimensions of the subspaces in Θ_T^{L0}. In Table 6, two formulae for D_T^{L0} appear in (8.16); the first follows immediately from the second definition for Θ_T^{L0} above it, and the second follows from the identity (4.9).

The space spanned by $\Theta_O^{\text{fiber}} = \Theta_O \setminus \Theta_O^{L0}$ has dimension $D_O^{\text{fiber}} = D_O - D_O^{L0}$; see the formula (8.12). Recall that we already derived this another way as (8.3).

A particularly important prebasis is $\Theta^{\text{free}} = \Theta_O^{\text{fiber}} \cup \Theta_T^{L0}$. By Lemma 39, the subspaces in Θ^{free} are linearly independent, so $D^{\text{free}} = D_O^{\text{fiber}} + D_T^{L0}$, giving us the right-hand side of the formula (8.19). Recall that Theorem 40 uses this fact to prove that span $\Theta^{\text{free}} = \text{null} \, d\mu(\theta)$. We now prove the middle part of (8.19).

Lemma 43. *The dimension of* null $d\mu(\theta)$ *can also be written*

$$D^{\text{free}} = d_\theta - \sum_{k,i \in [0,L], k+1 \geq i} \omega_{Li} \, \omega_{k0}.$$

Proof. By Theorem 40 and the identity (4.9), we have

$$
\begin{aligned}
D^{\text{free}} &= d_\theta - \sum_{j=1}^{L} \text{rk } W_{L\sim j} \cdot \text{rk } W_{j-1\sim 0} + \sum_{j=1}^{L-1} \text{rk } W_{L\sim j} \cdot \text{rk } W_{j\sim 0} \\
&= d_\theta - \text{rk } W_{L\sim L} \cdot \text{rk } W_{L-1\sim 0} - \sum_{j=1}^{L-1} \text{rk } W_{L\sim j} \cdot (\text{rk } W_{j-1\sim 0} - \text{rk } W_{j\sim 0}) \\
&= d_\theta - \text{rk } W_{L\sim L} \cdot (\omega_{L0} + \omega_{L-1,0}) - \sum_{j=1}^{L-1} \text{rk } W_{L\sim j} \cdot \left(\sum_{k=j-1}^{L} \omega_{k0} - \sum_{k=j}^{L} \omega_{k0} \right) \\
&= d_\theta - \text{rk } W_{L\sim L} \cdot \omega_{L0} - \sum_{j=1}^{L} \text{rk } W_{L\sim j} \cdot \omega_{j-1,0} \\
&= d_\theta - \omega_{L0} \sum_{i=0}^{L} \omega_{Li} - \sum_{j=1}^{L} \omega_{j-1,0} \sum_{i=0}^{j} \omega_{Li} \\
&= d_\theta - \omega_{L0} \sum_{i=0}^{L} \omega_{Li} - \sum_{k=0}^{L-1} \omega_{k0} \sum_{i=0}^{k+1} \omega_{Li} \\
&= d_\theta - \sum_{k,i\in[0,L],k+1\geq i} \omega_{Li}\,\omega_{k0}.
\end{aligned}
$$

$\square$

We interpret the summation in Lemma 43 as a count of the number of pairs of intervals of the form $([i,L],[0,k])$ that could undergo a connecting or swapping move. Any such move would change the value of $\mu(\theta)$, so it corresponds to a dimension of weight space that is not in the nullspace of the differential map.

Now consider Θ^{fiber}. Like Θ^{free}, its subspaces are linearly independent (Lemma 39). Recall from Section 8.4 that Θ^{fiber} can be constructed from Θ^{free} by removing some subspaces of the form ϕ_{lkjih}, each having dimension $\omega_{li}\,\omega_{kh}$, and replacing each such ϕ_{lkjih} with τ_{lkjih}, also having dimension $\omega_{li}\,\omega_{kh}$. Hence the sum of the dimensions of the subspaces does not change; $D^{\text{fiber}} = D^{\text{free}}$.

Recall from Section 7.4 the set of subspaces that specify connecting and swapping moves, $\Theta^{\text{comb}}_{\text{O}} = \{\phi_{lkjih} \neq \{\mathbf{0}\} : L \geq l \geq k+1 \geq j \geq i > h \geq 0\}$. To derive the dimension of the space spanned by $\Theta^{\text{comb}}_{\text{O}}$, we use the identities (4.8) and (4.9) and the fact that in the first summation below, the term $\omega_{li}\,\omega_{kh}$ appears $k - i + 2$ times—once for each $j \in [i, k+1]$.

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

Now consider $\Theta^{\text{comb}}_{\text{T}}$. As $\Theta^{\text{comb}}_{\text{T}} \subseteq \Theta^{\text{stratum}}$, the subspaces in $\Theta^{\text{comb}}_{\text{T}}$ are linearly independent (by Lemma 39). We determine $D^{\text{comb}}_{\text{T}} = \dim \text{span}\,\Theta^{\text{comb}}_{\text{T}}$ almost as we just determined $D^{\text{comb}}_{\text{O}}$, but the indices have the constraint $k \geq j$ (rather than $k + 1 \geq j$). Hence the term $\omega_{li}\,\omega_{kh}$ appears $k - i + 1$ times—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 D^{stratum}, we will use the difference between $D_{\mathrm{O}}^{\mathrm{comb}}$ and $D_{\mathrm{T}}^{\mathrm{comb}}$, which is

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

To derive D^{stratum} we must address the fact that Θ_{O}^{L0} and $\Theta_{\mathrm{O}}^{\mathrm{comb}}$ are not disjoint, nor are Θ_{T}^{L0} and $\Theta_{\mathrm{T}}^{\mathrm{comb}}$. Hence we define $\Theta_{\mathrm{O}}^{L0,\neg\mathrm{comb}} = \Theta_{\mathrm{O}}^{L0} \setminus \Theta_{\mathrm{O}}^{\mathrm{comb}} = \{\phi_{Lkji0} \in \Theta_{\mathrm{O}} : L = k \text{ or } i = 0\}$, a set of one-matrix subspaces representing moves off the fiber that are not connecting or swapping moves; that is, moves that change $\mu(\theta)$ but do not increase the rank of any subsequence matrix. With the identity (4.9), we find that the subspace span $\Theta_{\mathrm{O}}^{L0,\neg\mathrm{comb}}$ has dimension

$$
\begin{aligned}
D_{\mathrm{O}}^{L0,\neg\mathrm{comb}} &= \sum_{\phi_{Lkji0} \in \Theta_{\mathrm{O}}^{L0,\neg\mathrm{comb}}} \omega_{Li}\, \omega_{k0} = \sum_{\phi_{LLji0} \in \Theta_{\mathrm{O}}} \omega_{Li}\, \omega_{L0} + \sum_{\phi_{Lkj00} \in \Theta_{\mathrm{O}}} \omega_{L0}\, \omega_{k0} - \sum_{\phi_{LLj00} \in \Theta_{\mathrm{O}}} \omega_{L0}\, \omega_{L0} \\
&= \omega_{L0} \sum_{j=1}^{L} \left(\sum_{i=0}^{j} \omega_{Li} + \sum_{k=j-1}^{L} \omega_{k0} - \omega_{L0} \right) \\
&= \mathrm{rk}\, W \cdot \sum_{j=1}^{L} \left(\mathrm{rk}\, W_{L \sim j} + \mathrm{rk}\, W_{j-1 \sim 0} - \mathrm{rk}\, W \right),
\end{aligned}
$$

giving us the formula (8.14). Analogously, let $\Theta_{\mathrm{T}}^{L0,\neg\mathrm{comb}} = \Theta_{\mathrm{T}}^{L0} \setminus \Theta_{\mathrm{T}}^{\mathrm{comb}} = \{\tau_{Lkji0} \in \Theta_{\mathrm{T}} : L = k \text{ or } i = 0\}$. The space spanned by $\Theta_{\mathrm{T}}^{L0,\neg\mathrm{comb}}$ has dimension

$$
\begin{aligned}
D_{\mathrm{T}}^{L0,\neg\mathrm{comb}} &= \sum_{\tau_{Lkji0} \in \Theta_{\mathrm{T}}^{L0,\neg\mathrm{comb}}} \omega_{Li}\, \omega_{k0} = \sum_{\tau_{LLji0} \in \Theta_{\mathrm{T}}} \omega_{Li}\, \omega_{L0} + \sum_{\tau_{Lkj00} \in \Theta_{\mathrm{T}}} \omega_{L0}\, \omega_{k0} - \sum_{\tau_{LLj00} \in \Theta_{\mathrm{T}}} \omega_{L0}\, \omega_{L0} \\
&= \omega_{L0} \sum_{j=1}^{L-1} \left(\sum_{i=0}^{j} \omega_{Li} + \sum_{k=j}^{L} \omega_{k0} - \omega_{L0} \right) \\
&= \mathrm{rk}\, W \cdot \sum_{j=1}^{L-1} \left(\mathrm{rk}\, W_{L \sim j} + \mathrm{rk}\, W_{j \sim 0} - \mathrm{rk}\, W \right),
\end{aligned}
$$

giving us the formula (8.18). To derive D^{stratum}, we will use the difference between $D_{\mathrm{O}}^{L0,\neg\mathrm{comb}}$ and $D_{\mathrm{T}}^{L0,\neg\mathrm{comb}}$,

$$
\begin{aligned}
D_{\mathrm{O}}^{L0,\neg\mathrm{comb}} - D_{\mathrm{T}}^{L0,\neg\mathrm{comb}} &= \mathrm{rk}\, W \cdot (\mathrm{rk}\, W_{L \sim L} + \mathrm{rk}\, W_{0 \sim 0} - \mathrm{rk}\, W) \\
&= \mathrm{rk}\, W \cdot (d_L + d_0 - \mathrm{rk}\, W).
\end{aligned}
$$

At last we can derive D^{stratum}.

Lemma 44. *The dimension D^{stratum} of the subspace spanned by the stratum prebasis $\Theta^{\mathrm{stratum}}$ is specified by the formula (8.20).*

Proof. Recall that the subspaces in Θ^{stratum} are linearly independent, as are the subspaces in Θ_O; that $\Theta_O^{\text{comb}} \subseteq \Theta_O$ and $\Theta_O^{L0} \subseteq \Theta_O$; that $\Theta_O^{L0,\neg\text{comb}} = \Theta_O^{L0} \setminus \Theta_O^{\text{comb}}$ and $\Theta_T^{L0,\neg\text{comb}} = \Theta_T^{L0} \setminus \Theta_T^{\text{comb}}$; and that Θ_O is disjoint from $\Theta_T^{\text{comb}} \cup \Theta_T^{L0}$. Therefore,

$$
\begin{aligned}
D^{\text{stratum}} \;&=\; \dim \operatorname{span} \Theta^{\text{stratum}} \\
&=\; \dim \operatorname{span} \left(\left(\Theta_O \setminus \Theta_O^{\text{comb}} \setminus \Theta_O^{L0} \right) \cup \Theta_T^{\text{comb}} \cup \Theta_T^{L0} \right) \\
&=\; \dim \operatorname{span} \left(\left(\Theta_O \setminus \Theta_O^{\text{comb}} \setminus \Theta_O^{L0,\neg\text{comb}} \right) \cup \Theta_T^{\text{comb}} \cup \Theta_T^{L0,\neg\text{comb}} \right) \\
&=\; D_O - D_O^{\text{comb}} - D_O^{L0,\neg\text{comb}} + D_T^{\text{comb}} + D_T^{L0,\neg\text{comb}} \\
&=\; D_O - \left(D_O^{L0,\neg\text{comb}} - D_T^{L0,\neg\text{comb}} \right) - \left(D_O^{\text{comb}} - D_T^{\text{comb}} \right) \\
&=\; d_\theta - \operatorname{rk} W \cdot (d_L + d_0 - \operatorname{rk} W) - \sum_{L \geq k+1 \geq i > 0} \beta_{k+1,i,i}\, \alpha_{k,k,i-1}.
\end{aligned}
$$

$\square$

In Section 9.4, we will confirm that D^{stratum} is also the dimension of the stratum that contains θ, by showing that the subspace normal to that stratum at θ has dimension (at least) $d_\theta - D^{\text{stratum}}$.

9   The Normal Space, the Rowspace of d$\mu(\theta)$, and Prebases that Span Them

This section is devoted to determining two subspaces of the weight space $\mathbb{R}^{d_\theta}$: the subspace perpendicular to a stratum S at a weight vector $\theta \in S$, called the *normal space of S at θ* and written $N_\theta S$, and the rowspace of the differential map at θ, written row d$\mu(\theta)$. These two subspaces are the orthogonal complements of $T_\theta S$ and null d$\mu(\theta)$, respectively. The rowspace is of particular interest because gradient descent directions are usually chosen from row d$\mu(\theta)$.

One could say that by deriving $T_\theta S$ and null d$\mu(\theta)$ in Section 8, we have already derived $N_\theta S$ and row d$\mu(\theta)$, but in fact we have not finished deriving $T_\theta S$. In Sections 8.4 and 8.5, we specified two sets of subspaces, Θ^{free} and Θ^{stratum}, and we showed that null d$\mu(\theta) = \operatorname{span} \Theta^{\text{free}}$ and $T_\theta S \supseteq \operatorname{span} \Theta^{\text{stratum}}$. In Section 9.4 we will complete the proof that $T_\theta S = \operatorname{span} \Theta^{\text{stratum}}$, and thereby complete the proof that S and $T_\theta S$ both have dimension $D^{\text{stratum}} = \dim \operatorname{span} \Theta^{\text{stratum}}$, specified by the formula (8.20) in Table 6.

A second motivation for deriving row d$\mu(\theta)$ and $N_\theta S$ directly is to specify two prebases, Ψ^{free} and Ψ^{stratum}, that span row d$\mu(\theta)$ and $N_\theta S$. The number of basis vectors needed to express row d$\mu(\theta)$ and $N_\theta S$ is substantially smaller than the number of basis vectors needed to express null d$\mu(\theta)$ and $T_\theta S$. Hence, some computations are more easily done with Ψ^{free} and Ψ^{stratum} than with Θ^{free} and Θ^{stratum}.

Our conclusions are that row d$\mu(\theta) = \operatorname{span} \Psi^{\text{free}}$ and $N_\theta S = \operatorname{span} \Psi^{\text{stratum}}$ where the *freedom normal prebasis* Ψ^{free} and the *stratum normal prebasis* Ψ^{stratum} are defined to be

$$
\begin{aligned}
\Psi^{\text{free}} \;&=\; \{\psi_{Lki0} \neq \{\mathbf{0}\} : k, i \in [0, L] \text{ and } k + 1 \geq i\} \quad \text{and} & (9.1) \\
\Psi^{\text{stratum}} \;&=\; \Psi^{\text{free}} \cup \{\psi_{lkih} \neq \{\mathbf{0}\} : L \geq l \geq k + 1 \geq i > h \geq 0\}, & (9.2)
\end{aligned}
$$

and each ψ_{lkih} satisfying $L \geq l \geq i \geq 0$, $L \geq k \geq h \geq 0$, and $l > h$ is a subspace of $\mathbb{R}^{d_\theta}$ of the form

$$
\begin{aligned}
\psi_{lkih} \;&=\; \{(X_L, X_{L-1}, \ldots, X_1) : M \in b_{lli} \otimes a_{khh}\} \quad \text{where} & (9.3) \\
X_j \;&=\; \begin{cases} W_{l\sim j}^\top M W_{j-1\sim h}^\top, & j \in [i, k+1], \\ 0, & j \notin [i, k+1], \end{cases}
\end{aligned}
$$

and the prebasis subspaces b_{lli} and a_{khh} are defined in Section 4.3. For example, in the two-matrix case ($L = 2$),

$$
\begin{aligned}
\Psi^{\text{free}} &= \{\psi_{2000}, \psi_{2010}, \psi_{2100}, \psi_{2110}, \psi_{2120}, \psi_{2200}, \psi_{2210}, \psi_{2220}\} \quad \text{and} \\
\Psi^{\text{stratum}} &= \{\psi_{1010}, \psi_{2000}, \psi_{2010}, \psi_{2100}, \psi_{2110}, \psi_{2120}, \psi_{2121}, \psi_{2200}, \psi_{2210}, \psi_{2220}\}.
\end{aligned}
$$

Note that for $k + 1 \geq i$, the dimension of the subspace ψ_{lkih} is $\dim b_{lli} \cdot \dim a_{khh} = \omega_{li}\,\omega_{kh}$ and we can easily construct a basis for ψ_{lkih} given ω_{li} basis vectors for b_{lli} and ω_{kh} basis vectors for a_{khh}.

Recall from Theorem 42 that we establish that $T_\theta S = \text{span}\,\Theta^{\text{stratum}}$ by the following logic. We know from Lemmas 25 and 37 that $T_\theta S \supseteq \text{span}\,\Theta^{\text{stratum}}$. In Section 9.4, we will see that $N_\theta S \supseteq \text{span}\,\Psi^{\text{stratum}}$. (For both $T_\theta S$ and $N_\theta S$, it is easier to find all the subspaces they include than it is to verify that we have found enough subspaces!) We will count the dimensions of the subspaces in the prebasis Ψ^{stratum} and see that $\dim \text{span}\,\Theta^{\text{stratum}} + \dim \text{span}\,\Psi^{\text{stratum}} = d_\theta$. It follows that $T_\theta S = \text{span}\,\Theta^{\text{stratum}}$ and $N_\theta S = \text{span}\,\Psi^{\text{stratum}}$. From the former it follows that the dimension of S is D^{stratum}.

Sections 9.1–9.4 can be safely skipped by readers who don't want to know how these results were obtained.

9.1 The Rowspace of the Differential Map

Here we write an explicit expression for $\text{row}\,d\mu(\theta)$, the rowspace of the differential map. The differential map $d\mu(\theta)$ is a linear map from a weight displacement $\Delta\theta \in \mathbb{R}^{d_\theta}$ to a $d_L \times d_0$ matrix ΔW. We could represent $d\mu(\theta)$ as $(d_L d_0) \times d_\theta$ matrix, and apply to $d\mu(\theta)$ the same basic ideas from linear algebra that apply to any matrix, such as the nullspace, the columnspace (also known as the image), the rowspace, the left nullspace, and the rank. See Appendix C for a more detailed explanation.

Just like a matrix, the linear map $d\mu(\theta)$ has a transpose, which we write as $d\mu^\top(\theta)$. Recall from (8.1) that $d\mu(\theta)(\Delta\theta) = \sum_{j=1}^{L} W_{L\sim j}\Delta W_j W_{j-1\sim 0}$, where $\Delta\theta = (\Delta W_L, \Delta W_{L-1}, \ldots, \Delta W_1)$. Clearly this transformation is linear in $\Delta\theta$. It isn't written as a matrix-vector multiplication $M\Delta\theta$, but it could be. We can write the result of applying the transpose of $d\mu(\theta)$ (analogous to $M^\top$) to a matrix ΔW as

$$
d\mu^\top(\theta)(\Delta W) = (\Delta W W_{L-1\sim 0}^\top, \ldots, W_{L\sim j}^\top \Delta W W_{j-1\sim 0}^\top, \ldots, W_{L\sim 1}^\top \Delta W). \tag{9.4}
$$

Now we can write $\text{row}\,d\mu(\theta) = d\mu^\top(\theta)(\mathbb{R}^{d_L \times d_0})$, which is also called the *image* of $d\mu^\top(\theta)$ because it is the set found by applying $d\mu^\top(\theta)$ to every point in the domain $\mathbb{R}^{d_L \times d_0}$. In more detail,

$$
\begin{aligned}
\text{row}\,d\mu(\theta) &= \{(X_L, X_{L-1}, \ldots, X_1) : M \in \mathbb{R}^{d_L \times d_0}\} \quad \text{where} \\
X_j &= W_{L\sim j}^\top M W_{j-1\sim 0}^\top.
\end{aligned} \tag{9.5}
$$

For example, in the three-matrix case ($L = 3$),

$$
\text{row}\,d\mu(\theta) = \{(M W_1^\top W_2^\top, W_3^\top M W_1^\top, W_2^\top W_3^\top M) : M \in \mathbb{R}^{d_3 \times d_0}\}.
$$

This is the most explicit expression we will write for $\text{row}\,d\mu(\theta)$, and it is interesting to compare it with (8.8), our explicit expression for $\text{null}\,d\mu(\theta)$. As an enlightening exercise for the reader, we suggest verifying that any vector in (9.5) is orthogonal to any vector in (8.8). One unfortunate thing (9.5) and (8.8) have in common is that they reveal very little about the dimension of either subspace, nor about how to construct a basis for either subspace.

9.2   A Prebasis for the Rowspace of the Differential Map

Here we confirm that the rowspace of the differential map is $\operatorname{row} d\mu(\theta) = \operatorname{span} \Psi^{\text{free}}$, where Ψ^{free} is the freedom normal prebasis, specified by (9.1).

To begin, we decompose $\mathbb{R}^{d_L \times d_0}$ into a prebasis (a direct sum decomposition of $\mathbb{R}^{d_L \times d_0}$), then apply $d\mu^\top(\theta)$ to each subspace in the prebasis. We will define the subspaces in this prebasis with some extra generality that we will need in Section 9.4. The prebasis will contain some of the *subsequence subspaces*

$$e_{lkyxih} = b_{lyi} \otimes a_{kxh}, \qquad L \geq l \geq y \geq i \geq 0, L \geq k \geq x \geq h \geq 0, \text{ and } y > x. \tag{9.6}$$

Our results below require that all the subspaces of the form b_{lyi} and a_{kxh} are flow prebasis subspaces, as described in Section 4.4 and Theorem 12. The dimension of e_{lkyxih} is $\dim b_{lyi} \cdot \dim a_{kxh} = \omega_{li} \omega_{kh}$.

We group these subspaces into *subsequence prebases*. Given $y > x$, we define a subsequence prebasis that spans $\mathbb{R}^{d_y \times d_x}$.

$$\mathcal{E}_{yx} = \{e_{lkyxih} \neq \{0\} : l \in [y, L], k \in [x, L], i \in [0, y], h \in [0, x]\}. \tag{9.7}$$

This prebasis pairs every subspace $b_{lyi} \in \mathcal{B}_y$ with every subspace $a_{kxh} \in \mathcal{A}_x$. As $\mathcal{B}_y$ is a prebasis for $\mathbb{R}^{d_y}$ and $\mathcal{A}_x$ is a prebasis for $\mathbb{R}^{d_x}$ by Lemma 4, it is clear that $\mathcal{E}_{yx}$ is a prebasis for $\mathbb{R}^{d_y \times d_x}$. Hence we can uniquely represent a subsequence matrix $W_{y \sim x}$ as a vector sum of members of the subspaces in $\mathcal{E}_{yx}$.

In this section, we will use only $\mathcal{E}_{L0}$, which contains subspaces of the form e_{LkL0i0}. But in Section 9.4 we will use all the subsequence prebases.

Returning to the rowspace, we have

$$\operatorname{row} d\mu(\theta) = d\mu^\top(\theta)(\mathbb{R}^{d_L \times d_0}) = \operatorname{span} \{d\mu^\top(\theta)(e_{LkL0i0}) : e_{LkL0i0} \in \mathcal{E}_{L0}\}. \tag{9.8}$$

This motivates our defining, for all $k, i \in [0, L]$, the subspaces

$$\begin{aligned}
\psi_{Lki0} &= d\mu^\top(\theta)(e_{LkL0i0}) \\
&= \{(X_L, X_{L-1}, \ldots, X_1) : M \in e_{LkL0i0}\} \quad \text{where} \\
X_j &= W_{L \sim j}^\top M W_{j-1 \sim 0}^\top.
\end{aligned}$$

Then $\operatorname{row} d\mu(\theta) = \operatorname{span} \{\psi_{Lki0} : k, i \in [0, L]\}$. Next, we show that this subspace is $\operatorname{span} \Psi^{\text{free}}$.

Lemma 45. *For all $k, i \in [0, L]$, the subspace ψ_{Lki0} has the form (9.3). Moreover, if $k + 1 < i$ then $\psi_{Lki0} = \{0\}$. Moreover, assuming that the a- and b-spaces are flow subspaces, $X_j \in b_{Lji} \otimes a_{k,j-1,0}$ for $j \in [i, k+1]$.*

Proof. By definition, $e_{LkL0i0} = b_{LLi} \otimes a_{k00}$, so $X_j \in (W_{L \sim j}^\top b_{LLi}) \otimes (W_{j-1 \sim 0} a_{k00})$. Recall that for flow subspaces, $W_{L \sim j}^\top b_{LLi} = b_{Lji}$ if $j \geq i$; otherwise, $W_{L \sim j}^\top b_{LLi} = \{0\}$. Symmetrically, $W_{j-1 \sim 0} a_{k00} = a_{k,j-1,0}$ if $k \geq j - 1$; otherwise, $W_{j-1 \sim 0} a_{k00} = \{0\}$. Therefore, for any $M \in b_{LLi} \otimes a_{k00}$, $X_j = 0$ if $j \notin [i, k+1]$—hence ψ_{Lki0} has the form (9.3)—whereas $X_j \in b_{Lji} \otimes a_{k,j-1,0}$ if $j \in [i, k+1]$.

If $k + 1 < i$ then the range $[i, k+1]$ is empty, so $\psi_{Lki0} = \{0\}$.   □

Corollary 46. $\operatorname{row} d\mu(\theta) = \operatorname{span} \Psi^{\text{free}}$.

Proof. The identity (9.8) implies that $\operatorname{row} d\mu(\theta) = \operatorname{span} \{\psi_{Lki0} : k, i \in [0, L]\}$. By Lemma 45, if $k + 1 < i$ then $\psi_{Lki0} = \{0\}$. Hence the only difference between Ψ^{free} as defined by (9.1) and $\{\psi_{Lki0} : k, i \in [0, L]\}$ is that Ψ^{free} omits the trivial subspace $\{0\}$, which does not change the span. Therefore, $\operatorname{row} d\mu(\theta) = \operatorname{span} \Psi^{\text{free}}$.   □

In Section 9.4, we will see that the subspaces in Ψ^{free} are linearly independent (Lemma 50). Hence Ψ^{free} is a prebasis for row $d\mu(\theta)$. We will also see (in a more general context) that if $k + 1 \geq i$, then ψ_{Lki0} has the same dimension as e_{LkL0i0}, namely,

$$\dim \psi_{Lki0} = \dim e_{LkL0i0} = \omega_{Li}\, \omega_{k0}, \quad \text{for all } k, i \in [0, L] \text{ such that } k + 1 \geq i.$$

9.3 The Subspace Orthogonal to a Stratum

Here we write an explicit expression for $N_\theta S$, the (highest-dimensional) subspace normal to the stratum S at θ. To begin, recall from Section 5.1 that the stratum associated with a rank list $\underline{r}$ in the fiber of W is

$$S_{\underline{r}}^W = \mu^{-1}(W) \cap \bigcap_{L \geq y > x \geq 0} \text{WDM}_{\underline{r}}^{y \sim x}.$$

We use the abbreviation $S = S_{\underline{r}}^W$. Consider a weight vector $\theta \in S$ (thus θ has rank list $\underline{r}$). As S is a smooth manifold by Theorem 16, it has a well-defined normal space $N_\theta S$ and tangent space $T_\theta S$ at θ. As each weight-space determinantal manifold $\text{WDM}_{\underline{r}}^{y \sim x}$ is a smooth manifold too, we can write $T_\theta S \subseteq T_\theta \text{WDM}_{\underline{r}}^{y \sim x}$ for all y and x satisfying $L \geq y > x \geq 0$. The fiber $\mu^{-1}(W)$ is not a manifold and it might not have a tangent space at θ, but we can partly circumvent that by recalling that $T_\theta S \subseteq \text{null } d\mu(\theta)$ by Lemma 33. Hence

$$T_\theta S \subseteq \text{null } d\mu(\theta) \cap \bigcap_{L \geq y > x \geq 0} T_\theta \text{WDM}_{\underline{r}}^{y \sim x}. \tag{9.9}$$

Given subspaces satisfying $\sigma \subseteq \bigcap_j \sigma_j$, their orthogonal complements are related by $\sigma^\perp \supseteq \sum_j \sigma_j^\perp$. Hence

$$N_\theta S \supseteq \text{row } d\mu(\theta) + \sum_{L \geq y > x \geq 0} N_\theta \text{WDM}_{\underline{r}}^{y \sim x}, \tag{9.10}$$

where the sums are vector sums of subspaces of $\mathbb{R}^{d_\theta}$. We will show that the right-hand side of (9.10) is span Ψ^{stratum}, and thereby show that it is a subspace of dimension $d_\theta - D^{\text{stratum}}$ (see the forthcoming Lemma 51), so $N_\theta S$ has dimension at least $d_\theta - D^{\text{stratum}}$. We know that $T_\theta S$ has dimension at least D^{stratum} (Lemma 44), so we can replace the inclusion relations in (9.9) and (9.10) with equality (Theorem 52), thereby establishing that $N_\theta S = \text{span } \Psi^{\text{stratum}}$.

Recall that $\text{WDM}_{\underline{r}}^{y \sim x}$ is the set of weight vectors for which rk $W_{y \sim x} = r_{y \sim x}$. To determine $N_\theta \text{WDM}_{\underline{r}}^{y \sim x}$, the subspace normal to $\text{WDM}_{\underline{r}}^{y \sim x}$ at a weight vector θ, we introduce two polynomial functions. The first function, $\mu_{y \sim x}$, simply takes θ and returns the corresponding subsequence matrix $W_{y \sim x}$.

$$\mu_{y \sim x}(\theta) = W_y W_{y-1} \cdots W_{x+1}.$$

The second function, χ_r, maps a matrix M to a vector that is zero if and only if M has rank r or less. This vector lists the determinants of all the $(r + 1) \times (r + 1)$ minors of M. (Note that this vector might be *very* long—the number of minors can be exponential in the dimensions of M.) A $p \times q$ matrix M lies in the determinantal variety $\text{DV}_r^{p \times q}$ (has rank r or less) if and only if the determinant of every $(r+1) \times (r+1)$ minor of M is zero—that is, $\text{DV}_r^{p \times q} = \{M \in \mathbb{R}^{p \times q} : \chi_r(M) = \mathbf{0}\}$. M lies in the determinantal manifold $\text{DM}_r^{p \times q}$ (has rank exactly r) if and only if $\chi_r(M) = \mathbf{0}$ and $\chi_{r-1}(M) \neq \mathbf{0}$.

Each weight-space determinantal manifold is like a determinantal manifold, but the domain is weight space and r is the rank of a subsequence matrix. Given a rank list $\underline{r}$ and indices y and x, let $r = r_{y \sim x}$. Then

$$\text{WDM}_{\underline{r}}^{y \sim x} = \{\zeta \in \mathbb{R}^{d_\theta} : \chi_r(\mu_{y \sim x}(\zeta)) = \mathbf{0} \text{ and } \chi_{r-1}(\mu_{y \sim x}(\zeta)) \neq \mathbf{0}\}.$$

By the chain rule, the differential map of $\chi_r \circ \mu_{y \sim x}$ at θ is $d\chi_r(W_{y \sim x}) \circ d\mu_{y \sim x}(\theta)$, where $W_{y \sim x} = \mu_{y \sim x}(\theta)$. This differential map enables us to write the subspaces tangent and normal to $\mathrm{WDM}_r^{y \sim x}$ at θ,

$$
\begin{aligned}
T_\theta \mathrm{WDM}_{\underline{r}}^{y \sim x} &= \mathrm{null}\,(d\chi_r(W_{y \sim x}) \circ d\mu_{y \sim x}(\theta)) \quad \text{and} \\
N_\theta \mathrm{WDM}_{\underline{r}}^{y \sim x} &= \mathrm{row}\,(d\chi_r(W_{y \sim x}) \circ d\mu_{y \sim x}(\theta)) = d\mu_{y \sim x}^\top(\theta)(\mathrm{row}\, d\chi_r(W_{y \sim x})), \quad \text{where } r = r_{y \sim x}.
\end{aligned}
$$

To determine $N_\theta \mathrm{WDM}_{\underline{r}}^{y \sim x}$, we need to know $\mathrm{row}\, d\chi_r(W_{y \sim x})$. It is well known [9] that the subspaces tangent and normal to the determinantal variety $\mathrm{DV}_r^{p \times q}$ at a rank-r matrix M are

$$
\begin{aligned}
T_M \mathrm{DV}_r^{p \times q} &= \mathrm{col}\, M \otimes \mathbb{R}^q + \mathbb{R}^p \otimes \mathrm{row}\, M \quad \text{and} \\
N_M \mathrm{DV}_r^{p \times q} &= \mathrm{null}\, M^\top \otimes \mathrm{null}\, M.
\end{aligned}
$$

The latter implies that

$$
\begin{aligned}
\mathrm{row}\, d\chi_r(W_{y \sim x}) &= N_{W_{y \sim x}} \mathrm{DV}_r^{d_y \times d_x} = \mathrm{null}\, W_{y \sim x}^\top \otimes \mathrm{null}\, W_{y \sim x} \quad \text{and} \\
N_\theta \mathrm{WDM}_{\underline{r}}^{y \sim x} &= d\mu_{y \sim x}^\top(\theta)(\mathrm{null}\, W_{y \sim x}^\top \otimes \mathrm{null}\, W_{y \sim x}).
\end{aligned}
$$

Next, we need an expression for $d\mu_{y \sim x}^\top(\theta)$. Analogous to (8.1) and (9.4), by the product rule we have

$$
\begin{aligned}
d\mu_{y \sim x}(\theta)(\Delta\theta) &= \sum_{j=x+1}^{y} W_{y \sim j} \Delta W_j W_{j-1 \sim x}, \quad \text{and therefore} \\
d\mu_{y \sim x}^\top(\theta)(M) &= (X_L, X_{L-1}, \ldots, X_1) \quad \text{where} \\
X_j &= \begin{cases} W_{y \sim j}^\top M W_{j-1 \sim x}^\top, & j \in [x+1, y], \\ 0, & j \notin [x+1, y]. \end{cases}
\end{aligned}
\tag{9.11}
$$

Note that X_j is defined differently in (9.11) than it was in Section 9.1; it depends on the indices y and x. Now we can write the subspace normal to a weight-space determinantal manifold as

$$
N_\theta \mathrm{WDM}_{\underline{r}}^{y \sim x} = \{(X_L, X_{L-1}, \ldots, X_1) : M \in \mathrm{null}\, W_{y \sim x}^\top \otimes \mathrm{null}\, W_{y \sim x}\}.
\tag{9.12}
$$

Recalling (9.10) and (9.5), we can write the space normal to the stratum S as

$$
\begin{aligned}
N_\theta S &= \left\{(X_L^{L0}, X_{L-1}^{L0}, \ldots, X_1^{L0}) : M^{L0} \in \mathbb{R}^{d_L \times d_0}\right\} + \\
&\quad \sum_{L \geq y > x \geq 0} \left\{(X_L^{yx}, X_{L-1}^{yx}, \ldots, X_1^{yx}) : M^{yx} \in \mathrm{null}\, W_{y \sim x}^\top \otimes \mathrm{null}\, W_{y \sim x}\right\} \quad \text{where} \\
X_j^{yx} &= \begin{cases} W_{y \sim j}^\top M^{yx} W_{j-1 \sim x}^\top, & j \in [x+1, y], \\ 0, & j \notin [x+1, y]. \end{cases}
\end{aligned}
\tag{9.13}
$$

For example, in the two-matrix case ($L = 2$),

$$
N_\theta S = \{(MW_1^\top + M'', W_2^\top M + M') : M \in \mathbb{R}^{d_2 \times d_0}, M'' \in \mathrm{null}\, W_2^\top \otimes \mathrm{null}\, W_2, M' \in \mathrm{null}\, W_1^\top \otimes \mathrm{null}\, W_1\}.
$$

Note that we have not yet proven equality, only inclusion, as (9.13) rephrases (9.10); but as we will prove equality later we write (9.13) as an identity now. This is the most explicit expression we will write for $N_\theta S$; compare it with (8.9), our explicit expression for $T_\theta S$. The expressions (8.9) and (9.13), like (8.8) and (9.5), reveal little about the dimension of either subspace, nor about how to construct a basis for either subspace.

9.4 A Prebasis for the Subspace Orthogonal to a Stratum

Here we confirm that the subspace normal to the stratum S at θ is $N_\theta S = \text{span } \Psi^{\text{stratum}}$, where Ψ^{stratum} is the stratum normal prebasis, specified by (9.2). We also show that the dimension of $N_\theta S$ is $d_\theta - D^{\text{stratum}}$. As a consequence, we verify that our formula (8.20) for D^{stratum} is the dimension of S.

To begin, we use the flow subspaces $B_{y,y,x+1} = \text{null } W_{y\sim x}^\top$ and $A_{y-1,x,x} = \text{null } W_{y\sim x}$ from Section 4.2, their decomposition into prebasis subspaces b_{lyi} and a_{kxh} from Lemma 4, and the subsequence subspaces $e_{lkyxih} = b_{lyi} \otimes a_{kxh}$ defined by (9.6) in Section 9.2. We can write the subspace normal to the determinantal variety at $W_{y\sim x}$ as a vector sum of subsequence subspaces,

$$\text{null } W_{y\sim x}^\top \otimes \text{null } W_{y\sim x} = B_{y,y,x+1} \otimes A_{y-1,x,x} = \left(\sum_{l=y}^{L} \sum_{i=x+1}^{y} b_{lyi} \right) \otimes \left(\sum_{k=x}^{y-1} \sum_{h=0}^{x} a_{kxh} \right) = \sum_{l=y}^{L} \sum_{k=x}^{y-1} \sum_{i=x+1}^{y} \sum_{h=0}^{x} e_{lkyxih}.$$

The summands e_{lkyxih} are all members of the subsequence prebasis $\mathcal{E}_{yx}$, defined by (9.7) in Section 9.2. This expression implies that each subspace $e_{lkyxih} \in \mathcal{E}_{yx}$ is a subset of null $W_{y\sim x}^\top \otimes \text{null } W_{y\sim x}$ if $y > k$ and $i > x$. By substituting this identity into (9.12), we can write

$$N_\theta \text{WDM}_{\underline{r}}^{y\sim x} = \sum_{l=y}^{L} \sum_{k=x}^{y-1} \sum_{i=x+1}^{y} \sum_{h=0}^{x} \{(X_L, X_{L-1}, \dots, X_1) : M \in e_{lkyxih}\}.$$

This motivates our defining the subspaces

$$\psi_{lkyxih} = \{(X_L, X_{L-1}, \dots, X_1) : M \in e_{lkyxih}\}, \quad L \geq l \geq y \geq i \geq 0, L \geq k \geq x \geq h \geq 0, \text{ and } y > x,$$

so we can write $N_\theta WDM_{\underline{r}}^{y\sim x} = \sum_{l=y}^{L} \sum_{k=x}^{y-1} \sum_{i=x+1}^{y} \sum_{h=0}^{x} \psi_{lkyxih}$.

The following lemma elucidates the relationship between a matrix $M \in e_{lkyxih}$ and each corresponding matrix X_j in (9.11). One of its claims uses the flow bases described in Sections 4.6 and 4.7. Recall that each K_{lyi} is a $d_y \times \omega_{li}$ matrix whose columns form a basis for b_{lyi}, each J_{kxh} is a $d_x \times \omega_{kh}$ matrix whose columns form a basis for a_{kxh}, and the flow conditions state that $K_{lji} = W_{y\sim j}^\top K_{lyi}$ and $J_{k,j-1,h} = W_{j-1\sim x} J_{kxh}$.

Lemma 47. *For X_j defined by (9.11) where $M \in e_{lkyxih}$, $X_j = 0$ for each $j \notin [\max\{i, x+1\}, \min\{k+1, y\}]$, and the following claims hold for each X_j with $j \in [\max\{i, x+1\}, \min\{k+1, y\}]$.*

(a) *X_j has the same rank as M. In particular, $X_j = 0$ if and only if $M = 0$.*

(b) *$X_j \in e_{l,k,j,j-1,i,h} = b_{lji} \otimes a_{k,j-1,h}$.*

(c) *The identity (9.11), mapping M to X_j, is a bijection from e_{lkyxih} to $e_{l,k,j,j-1,i,h}$. Its inverse is the identity $M = K_{lyi} K_{lji}^+ X_j J_{k,j-1,h}^{+\top} J_{kxh}^\top$ where P^+ denotes the Moore–Penrose pseudoinverse of a matrix P and $P^{+\top}$ denotes the transpose of P^+.*

(d) *For every $z \in [\max\{i, x+1\}, \min\{k+1, y\}]$, $X_z = K_{lzi} K_{lji}^+ X_j J_{k,j-1,h}^{+\top} J_{k,z-1,h}^\top$. Hence X_j uniquely determines every other X_z in the range.*

(e) *For every matrix $X_j \in e_{l,k,j,j-1,i,h}$, ψ_{lkyxih} contains exactly one weight vector having that value of X_j.*

Proof. Following (9.11), consider the value of $X_j = W_{y\sim j}^\top M W_{j-1\sim x}^\top$ for some $j \in [x+1, y]$, given a matrix $M \in e_{lkyxih} = b_{lyi} \otimes a_{kxh}$. As the $\mathcal{B}_y$'s and $\mathcal{A}_x$'s are flow prebases, $W_{y\sim j}^\top b_{lyi} = b_{lji}$ if $j \geq i$; otherwise, $W_{y\sim j}^\top b_{lyi} = \{0\}$ (as $b_{lyi} \subseteq B_{lyi} \subseteq \text{null } W_{y\sim i-1}^\top$). Symmetrically, $W_{j-1\sim x} a_{kxh} = a_{k,j-1,h}$ if $k \geq j - 1$; otherwise,

$W_{j-1\sim x} a_{kxh} = \{\mathbf{0}\}$ (as $a_{kxh} \subseteq A_{kxh} \subseteq$ null $W_{k+1\sim x}$). Hence if $j \notin [\max\{i, x+1\}, \min\{k+1, y\}]$, then $X_j = 0$ as claimed. Otherwise, $X_j \in b_{lji} \otimes a_{k,j-1,h}$, confirming claim (b).

Any member of $b_{lyi} \otimes a_{kxh}$ can be written in the form $M = K_{lyi} C J_{kxh}^\top$ where C is an $\omega_{li} \times \omega_{kh}$ matrix of coefficients. (Each coefficient scales the outer product of one basis vector in K_{lyi} and one basis vector in J_{kxh}.) As K_{lyi} has rank ω_{li} and J_{kxh} has rank ω_{kh}, M has the same rank as C.

For any $j \in [\max\{i, x+1\}, \min\{k+1, y\}]$, $X_j = W_{y\sim j}^\top M W_{j-1\sim x}^\top = W_{y\sim j}^\top K_{lyi} C J_{kxh}^\top W_{j-1\sim x}^\top = K_{lji} C J_{k,j-1,h}^\top$. As K_{lji} has rank ω_{li} and $J_{k,j-1,h}$ has rank ω_{kh}, X_j has the same rank as C and we can recover C from X_j by writing $C = K_{lji}^+ X_j J_{k,j-1,h}^{+\top}$. Therefore rk $X_j =$ rk $C =$ rk M, confirming claim (a), and $M = K_{lyi} K_{lji}^+ X_j J_{k,j-1,h}^{+\top} J_{kxh}^\top$, confirming claim (c). Then for any $z \in [\max\{i, x+1\}, \min\{k+1, y\}]$,

$$
\begin{aligned}
X_z &= W_{y\sim z}^\top M W_{z-1\sim x}^\top \\
&= W_{y\sim z}^\top K_{lyi} K_{lji}^+ X_j J_{k,j-1,h}^{+\top} J_{kxh}^\top W_{z-1\sim x}^\top \\
&= K_{lzi} K_{lji}^+ X_j J_{k,j-1,h}^{+\top} J_{k,z-1,h}^\top,
\end{aligned}
$$

confirming claim (d).

Claim (c) implies that for every matrix $X_j \in e_{l,k,j-1,i,h}$, some weight vector in ψ_{lkyxih} has that value of X_j. Claim (d) implies that no other weight vector in ψ_{lkyxih} has that value of X_j, confirming claim (e). □

Lemma 48. *Consider a subspace ψ_{lkyxih} with valid indices (satisfying $L \geq l \geq y \geq i \geq 0$, $L \geq k \geq x \geq h \geq 0$, and $y > x$). If $k + 1 < i$ then $\psi_{lkyxih} = \{\mathbf{0}\}$; whereas if $k + 1 \geq i$ then ψ_{lkyxih} has the same dimension as e_{lkyxih}. That is,*

$$
\dim \psi_{lkyxih} = \dim e_{lkyxih} = \omega_{li} \omega_{kh}, \quad L \geq l \geq y \geq i \geq 0, L \geq k \geq x \geq h \geq 0, y > x, \text{ and } k+1 \geq i.
$$

Moreover, $\psi_{lkyxih} = \psi_{lky'x'ih}$ if $y > k$, $y' > k$, $i > x$, and $i > x'$ (assuming $\psi_{lky'x'ih}$ has valid indices too).

Proof. If $k + 1 < i$ then $\psi_{lkyxih} = \{\mathbf{0}\}$ by Lemma 47, because the range $[\max\{i, x+1\}, \min\{k+1, y\}]$ is empty. Whereas if $k + 1 \geq i$, then there exists at least one index $j \in [\max\{i, x+1\}, \min\{k+1, y\}]$, because $y \geq i$, $k \geq x$, and $y > x$ by assumption. As claim (c) of Lemma 47 applies to X_j, ψ_{lkyxih} has the same dimension as e_{lkyxih}, namely, $\omega_{li} \omega_{kh}$.

If $y > k$, $y' > k$, $i > x$, and $i > x'$ then $\psi_{lkyxih} = \psi_{lky'x'ih}$ because of the following three observations:

- $\min\{k + 1, y\} = k + 1 = \min\{k+1, y'\}$ and $\max\{i, x+1\} = i = \max\{i, x'+1\}$, so both subspaces have nonzero X_j only in the positions $j \in [i, k+1]$ by Lemma 47;

- by claim (d) of that lemma, each X_j in these positions determines the others through a formula independent of x and y; and

- by claims (e) and (b), both ψ_{lkyxih} and $\psi_{lky'x'ih}$ have a weight vector for every $X_j \in e_{l,k,j-1,i,h}$ and no weight vector with $X_j \notin e_{l,k,j-1,i,h}$.

□

We can write some subspaces with fewer indices by defining

$$
\psi_{lkih} = \psi_{lklhih}, \quad L \geq l \geq i \geq 0, L \geq k \geq h \geq 0, \text{ and } l > h.
$$

Corollary 49. *If ψ_{lkyxih} has valid indices satisfying $y > k$ and $i > x$, then $\psi_{lkyxih} = \psi_{lklhih} = \psi_{lkih}$.*

Proof. Valid indices imply that $l \geq y$ and $x \geq h$, so $l > k$ and $i > h$. By Lemma 48, $\psi_{lkyxih} = \psi_{lklhih}$. □

Recall that $N_\theta WDM_r^{y \sim x} = \sum_{l=y}^{L} \sum_{k=x}^{y-1} \sum_{i=x+1}^{y} \sum_{h=0}^{x} \psi_{lkyxih}$. The indices in the summations always satisfy $y > k$ and $i > x$, so by Corollary 49, we can shorten the indices to ψ_{lkih}. We can discard the subspaces with $k + 1 < i$, as they are all $\{\mathbf{0}\}$ by Lemma 48. Thus

$$
\begin{aligned}
N_\theta WDM_r^{y \sim x} &= \operatorname{span} \Psi_{y \sim x} \quad \text{where} \\
\Psi_{y \sim x} &= \{\psi_{lkih} \neq \{\mathbf{0}\} : l \in [y, L], k \in [x, y-1], i \in [x+1, y], h \in [0, x], \text{ and } k + 1 \geq i\}.
\end{aligned}
$$

We return to determining a prebasis for $N_\theta S$. It follows from (9.10) and Corollary 46 that $N_\theta S \supseteq \operatorname{span} \Psi^{\text{stratum}}$ where

$$
\begin{aligned}
\Psi^{\text{stratum}} &= \Psi^{\text{free}} \cup \bigcup_{L \geq y > x \geq 0} \Psi_{y \sim x} \\
&= \{\psi_{Lki0} \neq \{\mathbf{0}\} : k, i \in [0, L] \text{ and } k + 1 \geq i\} \cup \{\psi_{lkih} \neq \{\mathbf{0}\} : L \geq l \geq k + 1 \geq i > h \geq 0\}.
\end{aligned}
$$

This reiterates our definition (9.2) of Ψ^{stratum} from the beginning of Section 9. In this definition, some subspaces appear twice. For example, $\psi_{L,L-1,1,0}$ is in both the former and the latter set. The only subspaces unique to the first set are those of the form ψ_{LLi0} or ψ_{Lk00}. We can rewrite Ψ^{stratum} so that each subspace appears in exactly one set, which will help us count the dimension of $\operatorname{span} \Psi^{\text{stratum}}$.

$$
\begin{aligned}
\Psi^{\text{stratum}} &= \{\psi_{LLi0} \neq \{\mathbf{0}\} : i \in [0, L]\} \cup \{\psi_{Lk00} \neq \{\mathbf{0}\} : k \in [0, L-1]\} \cup \\
&\quad \{\psi_{lkih} \neq \{\mathbf{0}\} : L \geq l \geq k + 1 \geq i > h \geq 0\}.
\end{aligned}
\tag{9.14}
$$

The next two lemmas prove the linear independence of Ψ^{stratum} and derive the dimension of $\operatorname{span} \Psi^{\text{stratum}}$.

Lemma 50. *The subspaces in Ψ^{stratum} are linearly independent.*

Proof. Suppose for the sake of contradiction that Ψ^{stratum} is not linearly independent. Then there is a subspace $\psi_{lkih} \in \Psi^{\text{stratum}}$ and a nonzero weight vector $\zeta \in \psi_{lkih}$ such that ζ is a linear combination of weight vectors taken from the other subspaces in Ψ^{stratum}. Given a weight vector $\xi = (X_L, X_{L-1}, \ldots, X_1)$, let $M_j(\xi) = X_j$. Let j be an index such that $M_j(\zeta) \neq 0$. By claim (b) of Lemma 47, $M_j(\zeta) \in e_{l,k,j,j-1,i,h}$.

Given a subspace σ of weight vectors, let $M_j(\sigma) = \{M_j(\xi) : \xi \in \sigma\}$. By claims (b) and (e) of Lemma 47, for every subspace $\psi_{vuts} \in \Psi^{\text{stratum}}$, $M_j(\psi_{vuts}) = e_{v,u,j,j-1,t,s}$ or $M_j(\psi_{vuts}) = \{0\}$. Hence, distinct members of Ψ^{stratum} are mapped to $\{0\}$ or to distinct members of the prebasis $\mathcal{E}_{j,j-1}$, whose members are linearly independent. Therefore, $M_j(\zeta) \in e_{l,k,j,j-1,i,h}$ cannot be written as a linear combination of matrices taken from subspaces in $\mathcal{E}_{j,j-1} \setminus \{e_{l,k,j,j-1,i,h}\}$, so ζ cannot be written as a linear combination of weight vectors taken from $\Psi^{\text{stratum}} \setminus \{\psi_{lkih}\}$, a contradiction. It follows that the subspaces in Ψ^{stratum} are linearly independent. □

Lemma 51. *The dimension of* $\operatorname{span} \Psi^{\text{stratum}}$ *is $d_\theta - D^{\text{stratum}}$.*

Proof. The subspaces in Ψ^{stratum} are linearly independent by Lemma 50, so the dimension of $\operatorname{span} \Psi^{\text{stratum}}$ is the sum of the dimensions of the subspaces in Ψ^{stratum}, as written in (9.14). Recalling the formulae (4.1),

(4.3), and (4.8), this sum is

$$
\begin{aligned}
\sum_{\psi_{lkih} \in \Psi^{\text{stratum}}} \omega_{li}\, \omega_{kh} &= \sum_{i=0}^{L} \omega_{Li}\, \omega_{L0} + \sum_{k=0}^{L-1} \omega_{L0}\, \omega_{k0} + \sum_{L \geq k+1 \geq i > 0} \sum_{l=k+1}^{L} \sum_{h=0}^{i-1} \omega_{li}\, \omega_{kh} \\
&= \omega_{L0}\left(\sum_{i=0}^{L} \omega_{Li} + \sum_{k=0}^{L} \omega_{k0} - \omega_{L0} \right) + \sum_{L \geq k+1 \geq i > 0} \left(\sum_{l=k+1}^{L} \omega_{li} \right)\left(\sum_{h=0}^{i-1} \omega_{kh} \right) \\
&= \operatorname{rk} W \cdot (d_L + d_0 - \operatorname{rk} W) + \sum_{L \geq k+1 \geq i > 0} \beta_{k+1,i,i}\, \alpha_{k,k,i-1} \\
&= d_\theta - D^{\text{stratum}}.
\end{aligned}
$$

□

At last we complete our proof of the identities of $T_\theta S$ and $N_\theta S$, and our proof that the dimension of a stratum in the rank stratification is D^{stratum}.

Theorem 52. *Consider a matrix W, its fiber $\mu^{-1}(W)$, and the stratum $S = S_{\underline{r}}^{W}$ in W's fiber with rank list $\underline{r}$.*

- *The dimension of S is D^{stratum}.*
- *Θ^{stratum} is a prebasis for $T_\theta S$. (In particular, $T_\theta S = \operatorname{span} \Theta^{\text{stratum}}$.)*
- *Ψ^{stratum} is a prebasis for $N_\theta S$. (In particular, $N_\theta S = \operatorname{span} \Psi^{\text{stratum}}$.)*

Proof. By Lemmas 25 and 37, $\operatorname{span} \Theta^{\text{stratum}} \subseteq T_\theta S$. In this section we see that $\operatorname{span} \Psi^{\text{stratum}} \subseteq N_\theta S$. In Section 8.7 we saw that $\operatorname{span} \Theta^{\text{stratum}}$ has dimension D^{stratum}. By Lemma 51, $\operatorname{span} \Psi^{\text{stratum}}$ has dimension $d_\theta - D^{\text{stratum}}$. As $N_\theta S$ is the orthogonal complement of $T_\theta S$, $\dim N_\theta S + \dim T_\theta S = d_\theta$. The sum of the dimensions of $\operatorname{span} \Theta^{\text{stratum}}$ and $\operatorname{span} \Psi^{\text{stratum}}$ is d_θ, so $\operatorname{span} \Theta^{\text{stratum}} = T_\theta S$ and $\operatorname{span} \Psi^{\text{stratum}} = N_\theta S$. Hence the dimension of $T_\theta S$ is D^{stratum} and thus the dimension of S is D^{stratum}. As Θ^{stratum} is linearly independent by Lemma 39, Θ^{stratum} is a prebasis for $T_\theta S$. As Ψ^{stratum} is linearly independent by Lemma 50, Ψ^{stratum} is a prebasis for $N_\theta S$. □

10 Computing the Stratum Dag

Here we describe an algorithm for generating the stratum dag that represents the rank stratification of a fiber $\mu^{-1}(W)$—for instance, the stratum dag in Figure 4—given the integers $d_0, d_1, \ldots, d_L$, and $\operatorname{rk} W$ as input.

First we specify what information is stored with each vertex and directed edge of the dag (but we avoid specifying how the directed graph data structure is implemented). Each dag vertex v represents a nonempty stratum $S_{\underline{r}}$ in the rank stratification. The vertex data structure stores four fields: the numbers $v.\text{dim}$, $v.\text{dof}$, and $v.\text{rdof}$ introduced in Section 3 (recall Figures 2, 3, and 4), and $v.\text{ranklist}$ stores a rank list $\underline{r}$. The dimension of $S_{\underline{r}}$ is $v.\text{dim} = D^{\text{stratum}}$ from formula (8.20); the number of degrees of freedom of motion on the fiber from a point on $S_{\underline{r}}$ (i.e., the dimension of $\operatorname{null} d\mu(\theta)$) is $v.\text{dof} = D^{\text{free}}$ from formula (8.19); and the number of rank-increasing degrees of freedom is $v.\text{rdof} = v.\text{dof} - v.\text{dim}$. Each dag edge e represents a rank-1 abstract combinatorial move. The edge data structure stores three fields: $e.\text{origin}$ and $e.\text{dest}$ are the origin and destination vertices of the directed edge, and $e.\text{label}$ is a 4-tuple that lists the indices (l, k, i, h) associated with the combinatorial move.
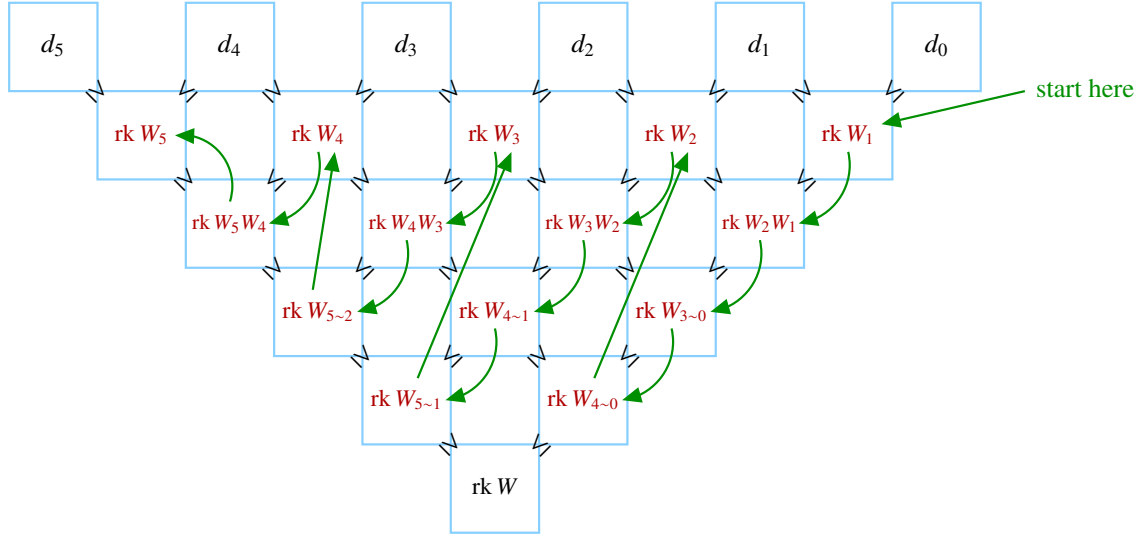
Figure 14: Green arrows indicate the order in which our algorithm for enumerating valid rank lists assigns ranks to the subsequence matrices. The d_j's and rk W are fixed (black text); only the ranks in red can vary from stratum to stratum. The $\geq$ and $\leq$ symbols indicate some of the constraints that the ranks must satisfy. There are additional constraints: each interval multiplicity $\omega_{ki} = \text{rk } W_{k\sim i} - \text{rk } W_{k\sim i-1} - \text{rk } W_{k+1\sim i} + \text{rk } W_{k+1\sim i-1}$ must be nonnegative (by the Frobenius rank inequality). For example, rk $W_2 \leq d_1 - \text{rk } W_1 + \text{rk } W_2 W_1$.

Our first task is to find a way to list all the strata in the rank stratification—that is, to enumerate all the valid rank lists that contain the specified values of the d_j's and rk W. This is not as easy as you might think. Section 10.1 provides an algorithm that creates one vertex in the stratum dag for each valid rank list, thus one vertex per stratum. (Appendix A shows that for every valid rank list $\underline{r}$, the stratum $S_{\underline{r}}$ is nonempty.)

Our second task is to create the edges of the stratum dag. These edges correspond to rank-1 abstract combinatorial moves. Section 10.2 gives an algorithm for identifying those moves and creating the edges and their labels.

10.1   Enumerating the Valid Rank Lists

Given specified values for the d_j's and rk W, this section describes a procedure to enumerate all the valid rank lists that include those values. We assume that rk $W \leq d_j$ for all $j \in [0, L]$; otherwise, no valid rank list includes those values and the fiber $\mu^{-1}(W)$ is the empty set.

Our procedure chooses the rank of one subsequence matrix, then recursively finds all the rank lists consistent with that choice. Then it repeats the process for every other valid choice of the matrix's rank. The order in which the ranks are chosen is depicted in Figure 14. Recall that for a specified fiber $\mu^{-1}(W)$, rk $W_{L\sim 0} = \text{rk } W$ and each rk $W_{j\sim j} = d_j$ are fixed. The other ranks rk $W_{k\sim i}$ are chosen in order of increasing i, and for ranks with equal i, in order of increasing k.

For each rank rk $W_{k\sim i}$ with $i = 0$, the procedure tries all values in the range

$$\text{rk } W_{k\sim 0} \in [\text{rk } W, \min\{\text{rk } W_{k-1\sim 0}, d_k\}], \qquad k \in [1, L-1].$$

For the other ranks ($i \neq 0$), the procedure tries all values in the range

$$\text{rk } W_{k\sim i} \in [\text{rk } W_{k\sim i-1}, \min\{\text{rk } W_{k-1\sim i} + \text{rk } W_{k\sim i-1} - \text{rk } W_{k-1\sim i-1}, d_k\}], \qquad L \geq k > i \geq 1.$$

The constraints that $\text{rk } W_{k\sim 0} \geq \text{rk } W$, $\text{rk } W_{k\sim i} \geq \text{rk } W_{k\sim i-1}$ for $i \geq 1$, and $\text{rk } W_{k\sim i} \leq d_k$ for all i are best understood from gazing at the $\geq$ and $\leq$ symbols in Figure 14. The constraint that $\text{rk } W_{k\sim i} \leq \text{rk } W_{k-1\sim i} + \text{rk } W_{k\sim i-1} - \text{rk } W_{k-1\sim i-1}$ comes from the requirement that the interval multiplicity $\omega_{k-1,i}$ is nonnegative.

The advantage of assigning ranks in this order is that, after we assign some of the ranks in the ranges indicated above, there is always at least one assignment of the remaining, unassigned ranks that satisfies the constraints and yields a valid rank list. We omit a proof, but a proof could proceed by showing that if we complete an incomplete rank list by always choosing the smallest rank in each range above (i.e., $\text{rk } W_{k\sim 0} \leftarrow \text{rk } W$ and $\text{rk } W_{k\sim i} \leftarrow \text{rk } W_{k\sim i-1}$ for $i \geq 1$), these choices satisfy the upper bound constraints and yield a valid rank list.

Figure 15 lists pseudocode for a procedure, ENUMERATEVERTICES, that enumerates the valid rank lists (and the strata in the rank stratification) for a fiber $\mu^{-1}(W)$. For each valid rank list $\underline{r}$, the algorithm creates a vertex in the stratum dag which represents both $\underline{r}$ and the stratum $S_{\underline{r}}$. ENUMERATEVERTICES creates no edges; see Section 10.2 for the procedure that creates them.

Our algorithm maintains a map from rank lists to (references to) graph vertices. This map is not part of the directed graph data structure, but it is used to find vertices in the graph data structure knowing only their rank lists, and it will be needed in Section 10.2 to help compute the edges of the dag. The procedure ADDTOMAP (called in Line 25) adds a rank list (the key) and a vertex (the target) to this map. There are several ways the map could be implemented. The simplest method is to store the vertices in a multidimensional array indexed by the ranks in the rank list. If we omit the fixed ranks (the d_j's and $\text{rk } W$), this array has $(L^2 + L)/2 - 1$ dimensions, and the index associated with $r_{k\sim i}$ can range from $\text{rk } W$ to $\min\{d_i, d_{i+1}, \ldots, d_k\}$. However, if L is large, most of the entries in this array are never used, as they correspond to invalid rank lists, and the array might be too large to store. In that case, it makes more sense to use a hash table to implement the map.

10.2   Computing the Edges of the Stratum Dag

In Section 3.1, we introduced the convention that the stratum dag has a directed edge $(S_{\underline{r}}, S_{\underline{s}})$ if there exists a rank-1 abstract combinatorial move that transforms the rank list $\underline{r}$ to the rank list $\underline{s}$. This idea is partly justified by Theorem 27, which establishes that, assuming $S_{\underline{r}}$ and $S_{\underline{s}}$ are both nonempty strata in the rank stratification of $\mu^{-1}(W)$, the following three statements are equivalent: $S_{\underline{r}} \subseteq \bar{S}_{\underline{s}}$; $\underline{r} \leq \underline{s}$; and there exists a sequence of rank-1 abstract connecting and swapping moves that proceed from $\underline{r}$ to $\underline{s}$, with all the intermediate rank lists being valid. The last statement is equivalent to there being a directed path from $S_{\underline{r}}$ to $S_{\underline{s}}$ in the stratum dag.

Figure 16 lists pseudocode for a procedure, ENUMERATEEDGES, that creates the directed edges of the stratum dag for a fiber $\mu^{-1}(W)$, connecting the vertices already created by the procedure ENUMERATEVERTICES. We consider one stratum $S_{\underline{r}}$ at a time, and enumerate all the rank-1 abstract combinatorial moves from $\underline{r}$; these correspond to directed edges out of $S_{\underline{r}}$ in the stratum dag. From $\underline{r}$, there is an abstract combinatorial move with index tuple (l, k, i, h) if $L \geq l \geq k + 1 \geq i > h \geq 0$, $\omega_{li} > 0$, and $\omega_{kh} > 0$ (see Section 7.5). Recall from Lemma 24 (and Section 7.4 and Table 5) that a rank-1 combinatorial move with these indices increases each rank $r_{y\sim x}$ by 1 for all $y \in [k + 1, l]$ and $x \in [h, i - 1]$. This tells us how to compute the rank list $\underline{s}$ reached by a combinatorial move, whereupon we can create a directed edge $(S_{\underline{r}}, S_{\underline{s}})$ in the dag.

Line 1 of ENUMERATEEDGES presumes that the directed graph data structure that represents the stratum dag permits us to easily loop through the complete set of graph vertices (which were originally created by ENUMERATEVERTICES). Line 15 uses the map maintained by Line 25 of ENUMERATEVERTICES to find the data structure representing a dag vertex, knowing only its rank list $\underline{s}$. Lines 16 and 17 add an edge to the graph data structure, labeled with the index tuple (l, k, i, h).

ENUMERATEVERTICES($L, d[], R$)

{ Enumerates all the valid rank lists that include specified values of $d_0, d_1, \ldots, d_L$ and rk $W = R$. }

{ Creates one dag vertex for each valid rank list. L is the number of matrices (i.e., layers of edges). }

1       $r \leftarrow$ rank list in which every rank $r_{k\sim i}$ is zero

2       $r_{L\sim0} \leftarrow R$ { in r, initialize $r_{L\sim0}$ to rk W }

3   **for** $j \leftarrow 0$ **to** L

4           $r_{j\sim j} \leftarrow d_j$

5   ENUMERATEVERTICESRECURSE($L, d[], r, 1, 0$)   { start the recursion with $r_{1\sim0}$, representing rk W_1 }

ENUMERATEVERTICESRECURSE($L, d[], r, k, i$)

{ Iterates through all the valid values of $r_{k\sim i}$, then recursively does the same for the remaining ranks. }

6   **if** $i = 0$

           { There is at least one valid rank list for each rk $W_{k\sim0} \in [\text{rk } W, \min\{\text{rk } W_{k-1\sim0}, d_k\}]$. }

7   low $\leftarrow r_{L\sim0}$ { note: $r_{L\sim0} = $ rk W }

8   high $\leftarrow \min\{r_{k-1\sim0}, d_k\}$

9   **else**   { $i \geq 1$ }

           { ... for each rk $W_{k\sim i} \in [\text{rk } W_{k\sim i-1}, \min\{\text{rk } W_{k-1\sim i} + \text{rk } W_{k\sim i-1} - \text{rk } W_{k-1\sim i-1}, d_k\}]$. }

10   low $\leftarrow r_{k\sim i-1}$

11   high $\leftarrow \min\{r_{k-1\sim i} + r_{k\sim i-1} - r_{k-1\sim i-1}, d_k\}$

12   **if** $k < L$ **and** ($k < L - 1$ **or** $i > 0$)

13   **for** $r_{k\sim i} \leftarrow$ low **to** high

14   ENUMERATEVERTICESRECURSE($L, d[], r, k + 1, i$)   { increment k }

15   **else if** $i < L - 1$

16   **for** $r_{k\sim i} \leftarrow$ low **to** high

17   ENUMERATEVERTICESRECURSE($L, d[], r, i + 2, i + 1$)   { increment i; restart k }

18   **else**   { $k = L$ and $i = L - 1$; base case of recursion }

19   **for** $r_{k\sim i} \leftarrow$ low **to** high

           { All the ranks in r are filled in now; create a new vertex to represent the stratum S_r. }

20              $v \leftarrow$ CREATEDAGVERTEX()   { creates new vertex in dag data structure }

21   v.ranklist $\leftarrow r$ { important: copy the rank list r; don't just store a reference to r }

           { v.dim is the dimension of the stratum S_r. }

22   v.dim $\leftarrow \sum_{j=1}^{L} d_j d_{j-1} - r_{L\sim0} \cdot (d_L + d_0 - r_{L\sim0}) - \sum_{L \geq k+1 \geq i > 0} (r_{k+1\sim i} - r_{k+1\sim i-1})(r_{k\sim i-1} - r_{k+1\sim i-1})$

           { v.dof is the number of degrees of freedom of motion on the fiber from a point on S_r. }

23   v.dof $\leftarrow \sum_{j=1}^{L} d_j d_{j-1} - \sum_{j=1}^{L} r_{L\sim j} \cdot r_{j-1\sim0} + \sum_{j=1}^{L-1} r_{L\sim j} \cdot r_{j\sim0}$

           { v.rdof is the number of rank-increasing degrees of freedom (moving off S_r on fiber). }

24   v.rdof $\leftarrow v$.dof $- v$.dim

           { Maintain a map from rank lists to dag vertices so the record v can be found later. }

25   ADDTOMAP(r, v)

Figure 15: Algorithm to enumerate all the strata in the rank stratification of a fiber $\mu^{-1}(W)$—equivalently, to enumerate all the valid rank lists that include specified values of $d_0, d_1, \ldots, d_L$, and rk W. For each such rank list r, the algorithm creates a vertex in the stratum dag to represent the stratum S_r. A dag vertex v stores four fields: the numbers v.dim, v.dof, and v.rdof introduced in Section 3, and v.ranklist stores a rank list r.

ENUMERATEEDGES(L)
{ Creates the edges of the stratum dag, connecting the vertices created by ENUMERATEVERTICES. }
{ L is the number of matrices (i.e., layers of neural network edges). }
1   **for** each vertex v in the stratum dag
2        $\underline{r} = v$.ranklist
         { Compute the interval multiplicities for $\underline{r}$. }
3   **for** $y \leftarrow 0$ **to** L
4   **for** $x \leftarrow 0$ **to** y
                 { For the next line, use the convention $r_{L+1\sim x} = r_{y\sim-1} = 0$. }
5                $\omega_{yx} \leftarrow r_{y\sim x} - r_{y\sim x-1} - r_{y+1\sim x} + r_{y+1\sim x-1}$
         { Identify all possible rank-1 abstract combinatorial moves from $\underline{r}$. }
6   **for** $l \leftarrow 1$ **to** L
7   **for** $k \leftarrow 0$ **to** $l - 1$
8   **for** $i \leftarrow 1$ **to** $k + 1$
9   **for** $h \leftarrow 0$ **to** $i - 1$
10   **if** $\omega_{li} > 0$ **and** $\omega_{kh} > 0$
                             { There is a move connecting or swapping intervals $[i, l]$ and $[h, k]$. }
11                           $\underline{s} \leftarrow \underline{r}$ { copy the rank list $\underline{r}$; don't just store a reference to $\underline{r}$ }
12   **for** $y \leftarrow k + 1$ **to** l
13   **for** $x \leftarrow h$ **to** $i - 1$
14                                   $s_{y\sim x} \leftarrow r_{y\sim x} + 1$
                             { Add directed edge representing $(S_{\underline{r}}, S_{\underline{s}})$ to dag data structure. }
15                           $w \leftarrow$ MAPRANKLISTTOVERTEX($\underline{s}$)
16                           $e \leftarrow$ CREATEDAGEDGE(v, w)   { sets e.origin to v and e.dest to w }
17   e.label $\leftarrow (l, k, i, h)$

Figure 16: Algorithm to create the edges of the stratum dag. The procedure MAPRANKLISTTOVERTEX invoked in Line 15 performs a lookup in the same map maintained by the procedure ADDTOMAP in Line 25 of ENUMERATEVERTICESRECURSE.

10.3   An Alternative Method of Enumerating the Valid Rank Lists

We briefly mention an alternative to the algorithm ENUMERATEVERTICES. Our alternative builds on the structure of ENUMERATEEDGES. Recall that every fiber's rank stratification has one minimal element: the stratum for which every rank (except the d_j's) is equal to rk W. Every other stratum in the rank stratification can be reached from the minimal stratum by a sequence of rank-1 combinatorial moves. Hence we can efficiently find all the vertices with a procedure that is essentially depth-first search, except that the graph is not explicitly stored as a data structure (until the search is complete; we create the data structure as we go).

Therefore, we can modify ENUMERATEEDGES to create all the dag vertices as well. Line 15 is modified to check whether a dag vertex exists to represent the rank list $\underline{s}$; if no such vertex exists, one is created and added to a queue of vertices whose outbound edges have not yet been created. Line 1 is modified to iterate through that queue. The algorithm begins by creating one dag vertex to represent the minimal rank list, then creating a queue that contains that vertex. As the algorithm runs, newly created vertices are added to the queue, and the loop in Line 1 is modified to repeatedly take a vertex off the queue and determine its outbound edges.

A   Valid Rank Lists and Valid Multisets of Intervals

Consider a fully-connected linear neural network specified by the unit layer sizes $d_0, d_1, \ldots, d_L$, with a corresponding weight space $\mathbb{R}^{d_\theta}$. Recall that a rank list $\underline{r}$ is *valid* if there is some weight vector $\theta \in \mathbb{R}^{d_\theta}$ that has rank list $\underline{r}$. (Note that this implies that $r_{j \sim j} = d_j$ for all $j \in [0, L]$.) Recall that a multiset of intervals, represented by interval multiplicities ω_{ki}, $L \geq k \geq i \geq 0$, is *valid* for this network if $d_j = \sum_{t=j}^{L} \sum_{s=0}^{j} \omega_{ts}$ for every $j \in [0, L]$. We also require that every ω_{ki} is nonnegative, but that's part of the definition of "multiset," not part of the definition of "valid."

The identities (4.9) and (4.10) map interval multiplicities to rank lists and vice versa. When the rank list and interval multiplicities are properties of a weight vector $\theta \in \mathbb{R}^{d_\theta}$, the two identities are correct by Lemma 10, so they must define a bijection between valid rank lists and multisets of intervals that arise from weight vectors. However, we did not define a multiset of intervals to be "valid" if it arises from a weight vector; we opted for a definition that is easier to check.

Here we justify that definition by showing that (4.9) maps every valid multiset of intervals to a valid rank list, and (4.10) maps every valid rank list to a valid multiset of intervals. Consider the latter first.

Lemma 53. *Let $\underline{r}$ be a valid rank list. Then the identity (4.10) yields a valid multiset of intervals (with validity judged according to the unit layer sizes d_j specified in $\underline{r}$).*

Proof. As $\underline{r}$ is valid, there exists a weight vector $\theta = (W_L, W_{L-1}, \ldots, W_1) \in \mathbb{R}^{d_\theta}$ with rank list $\underline{r}$. The identity (4.10) is $\omega_{ki} = \text{rk } W_{k \sim i} - \text{rk } W_{k \sim i-1} - \text{rk } W_{k+1 \sim i} + \text{rk } W_{k+1 \sim i-1}$ with the conventions that $W_{L+1} = 0$ and $W_0 = 0$, hence every interval multiplicity ω_{ki} is nonnegative by the Frobenius rank inequality (Sylvester's inequality if $k = i$). This establishes that the interval multiplicities specify a multiset.

The criterion for this multiset to be valid is that $d_j = \sum_{t=j}^{L} \sum_{s=0}^{j} \omega_{ts}$ for every $j \in [0, L]$, which is what Corollary 9 states. □

Showing that a valid multiset of intervals maps to a valid rank list is more interesting, because we have to exhibit a weight vector with the specified rank list. Fortunately, we constructed one in Section 4.6. Recall that for any valid multiset of intervals, Lemma 11 describes a canonical weight vector $\tilde{\theta} = (\tilde{I}_L, \tilde{I}_{L-1}, \ldots, \tilde{I}_1)$ that has the specified interval multiplicities. (Appendix B makes a few comments about the prebasis subspaces associated with $\tilde{\theta}$.) We will omit the details of verifying that, so the following proof is just a proof sketch. (There are times when a formal proof that a construction is correct would be more cryptic than simply inspecting the construction until you see why it's correct, and we think this is one of them. Figure 8 is the real proof.)

Lemma 54. *Given a network with specified unit layer sizes $d_0, d_1, \ldots, d_L$, suppose the interval multiplicities ω_{ki}, $L \geq k \geq i \geq 0$, represent a valid multiset of intervals. Let $\underline{r}$ be the rank list produced by the identity (4.9) from those interval multiplicities. Then $\underline{r}$ is a valid rank list.*

Proof. The canonical weight vector $\tilde{\theta}$ described by Lemma 11 has the specified interval multiplicities. By Lemma 10, the rank list $\underline{r}$ produced by identity (4.9) is the rank list of $\tilde{\theta}$. Hence $\underline{r}$ is valid. □

Thus both definitions of "valid" imply that there exists a weight vector $\theta \in \mathbb{R}^{d_\theta}$ with the specified rank list or interval multiplicities. But often we are interested in the question of whether there exists a suitable weight vector on a particular fiber $\mu^{-1}(W)$. For a specified rank list $\underline{r}$, this is possible only if $\text{rk } W = r_{L \sim 0}$; and for a specified multiset of intervals, it is possible only if $\text{rk } W = \omega_{L0}$ by (4.3). But if that constrain holds, then a

suitable weight vector $\theta \in \mu^{-1}(W)$ always exists, and we now show how to construct one. Our construction combines a singular value decomposition of W with the canonical weight vector $\tilde{\theta}$.

It is well known that W has a singular value decomposition $W = UDV^\top$ such that $U \in \mathbb{R}^{d_L \times d_L}$ and $V \in \mathbb{R}^{d_0 \times d_0}$ are both orthogonal matrices ($UU^\top = I = U^\top U$ and $VV^\top = I = V^\top V$) and $D \in \mathbb{R}^{d_L \times d_0}$ is a diagonal (not necessarily square) matrix, whose diagonal components are the singular values of W. Moreover, we can choose D so that the nonzero singular values come before the zeros on the diagonal. Of these singular values, ω_{L0} are nonzero and the rest are zero (because rk $W = \omega_{L0}$).

Recall that the product $\tilde{I} = \mu(\tilde{\theta}) = \tilde{I}_L \tilde{I}_{L-1} \cdots \tilde{I}_1$ is a diagonal $d_L \times d_0$ matrix whose first ω_{L0} diagonal components are 1's, and all the other components are 0's. Hence, the nonzero diagonal components of $\tilde{I}$ and D are in the same positions. Let D' be a diagonal $d_0 \times d_0$ matrix whose first ω_{L0} diagonal components are the same as D's (the nonzero singular values), and whose remaining diagonal components are 1's, making D' invertible. Then $\tilde{I}D' = D$ and

$$W = UDV^\top = U\tilde{I}D'V^\top = U\tilde{I}_L\tilde{I}_{L-1} \cdots \tilde{I}_2\tilde{I}_1 D'V^\top.$$

As U and $D'V^\top$ are invertible, the weight vector

$$\theta = (U\tilde{I}_L, \tilde{I}_{L-1}, \tilde{I}_{L-2}, \ldots, \tilde{I}_2, \tilde{I}_1 D'V^\top)$$

has the same rank list as $\tilde{\theta}$ and satisfies $\mu(\theta) = W$.

The following two lemmas use this construction to show that for either a valid rank list or a valid multiset of intervals, the fiber $\mu^{-1}(W)$ contains a point with that rank list or multiset of intervals if the rank of W is a match (i.e., rk $W = r_{L\sim0}$ or rk $W = \omega_{L0}$).

Lemma 55. *Let $\underline{r}$ be a valid rank list and let $W \in \mathbb{R}^{d_L \times d_0}$ be a matrix of rank $r_{L\sim0}$. Then there is a weight vector $\theta \in \mu^{-1}(W)$ with rank list $\underline{r}$.*

Proof. As $\underline{r}$ is valid, by definition there exists a weight vector $\theta_{\underline{r}} \in \mathbb{R}^{d_\theta}$ with rank list $\underline{r}$. Section 4.6 describes how to construct a canonical weight vector $\tilde{\theta} = \eta^{-1}(\theta_{\underline{r}})$ also having rank list $\underline{r}$, where η is defined in Section 5.2. As rk $W = r_{L\sim0}$, the procedure described above constructs a weight vector θ having the same rank list as $\tilde{\theta}$, namely $\underline{r}$, such that $\mu(\theta) = W$.   □

Lemma 56. *Suppose the interval multiplicities ω_{ki}, $L \geq k \geq i \geq 0$, represent a valid multiset of intervals for a network with specified unit layer sizes $d_0, d_1, \ldots, d_L$. Let $\underline{r}$ be the rank list produced by the identity (4.9) from those interval multiplicities. (Note that by the definition of "valid," $\underline{r}$ satisfies $r_{j\sim j} = d_j$ for $j \in [0, L]$.) Then for any matrix $W \in \mathbb{R}^{d_L \times d_0}$ of rank ω_{L0}, there is a weight vector $\theta \in \mu^{-1}(W)$ with rank list $\underline{r}$.*

Proof. The canonical weight vector $\tilde{\theta}$ described by Lemma 11 has the specified interval multiplicities. By Lemma 10, the rank list $\underline{r}$ produced by identity (4.9) is the rank list of $\tilde{\theta}$. As rk $W = \omega_{L0}$, the procedure described above constructs a weight vector θ having the same rank list as $\tilde{\theta}$, namely $\underline{r}$, such that $\mu(\theta) = W$.   □

B   The Standard Prebases

Recall from Section 4.3 that the *standard prebases* $\mathcal{A}_0, \ldots, \mathcal{A}_L, \mathcal{B}_0, \ldots, \mathcal{B}_L$ are defined by setting $a_{kji} = A_{kji} \cap (A_{k,j,i-1} + A_{k-1,j,i})^\perp$ and $b_{kji} = B_{kji} \cap (B_{k,j,i+1} + B_{k+1,j,i})^\perp$, which can also be written as (4.6) and (4.7). Although it is not generally possible to have all the subspaces in $\mathcal{A}_j$ be pairwise orthogonal, nor all the subspaces in $\mathcal{B}_j$, the subspaces in the standard prebasis are as close to orthogonal as possible, which is good

for the numerical stability of algorithms that use them. There are four exceptions: the subspaces in any one of $\mathcal{A}_0$, $\mathcal{A}_L$, $\mathcal{B}_0$, and $\mathcal{B}_L$ are pairwise orthogonal. (Subspaces that meet at oblique angles occur only in the hidden layers.) The standard prebases have the property that $\mathcal{A}_0 = \mathcal{B}_0$ and $\mathcal{A}_L = \mathcal{B}_L$, because for all $k, i \in [0, L]$,

$$a_{k00} = A_{k00} \cap A_{k-1,0,0}^{\perp} = \text{null } W_{k+1\sim 0} \cap \text{row } W_{k\sim 0} = B_{k+1,0,0}^{\perp} \cap B_{k00} = b_{k00} \quad \text{and}$$

$$b_{LLi} = B_{LLi} \cap B_{L,L,i+1}^{\perp} = \text{null } W_{L\sim i-1}^{\top} \cap \text{col } W_{L\sim i} = A_{L,L,i-1}^{\perp} \cap A_{LLi} = a_{LLi}.$$

However, in the hidden layers, $\mathcal{A}_j$ and $\mathcal{B}_j$ can be entirely different.

Unfortunately, the standard prebases are typically not flow prebases. But at the canonical weight vector $\tilde{\theta}$ (introduced in Section 4.6), the standard prebases *are* flow prebases. The standard prebases at $\tilde{\theta}$ have other special properties too. We say that a subspace of $\mathbb{R}^d$ is *axis-aligned* if it is spanned by some of the unit coordinate vectors of $\mathbb{R}^d$. Recall that the almost-identity matrices, including the factor matrices $\tilde{I}_j$ and the subsequence matrices $\tilde{I}_{y\sim x}$, have the property that every row and every column has at most one nonzero component, always a 1. Hence all subspaces of the forms row $\tilde{I}_{y\sim x}$, null $\tilde{I}_{y\sim x}$, col $\tilde{I}_{y\sim x}$, and null $\tilde{I}_{y\sim x}^{\top}$ are axis-aligned. Hence all subspaces of the forms A_{kji} and B_{kji} are axis-aligned. Hence in the standard prebases, all subspaces of the forms a_{kji} and b_{kji} are axis-aligned. As the subspaces in each standard prebasis $\mathcal{A}_j$ are linearly independent and axis-aligned, they are pairwise orthogonal. (Likewise for $\mathcal{B}_j$.)

C The Fundamental Subspaces of the Differential Map

The differential map $d\mu(\theta)$ is a linear operator that maps a weight displacement $\Delta\theta$ in a finite vector space $\mathbb{R}^{d_\theta}$ to a matrix $\Delta W = d\mu(\theta)(\Delta\theta)$ in another finite vector space $\mathbb{R}^{d_L \times d_0}$. One could simply write this linear map as a matrix—a fact that is obscured because the range of $d\mu(\theta)$ is a set of matrices. But if we imagine that the range of $d\mu(\theta)$ is $\mathbb{R}^{d_L d_0}$ instead of $\mathbb{R}^{d_L \times d_0}$, we can represent $d\mu(\theta)$ as a $(d_L d_0) \times d_\theta$ matrix M. Then we can apply basic ideas from linear algebra.

The four fundamental subspaces of M are its nullspace, its columnspace (also called its image), its rowspace, and its left nullspace, defined to be

$$\begin{aligned}
\text{null } M \quad &= \quad \{v \in \mathbb{R}^{d_\theta} : Mv = \mathbf{0}\}, \\
\text{col } M = \text{image } M \quad &= \quad \{Mv : v \in \mathbb{R}^{d_\theta}\}, \\
\text{row } M = \text{image } M^{\top} \quad &= \quad \{M^{\top}w : w \in \mathbb{R}^{d_L d_0}\}, \quad \text{and} \\
\text{null } M^{\top} \quad &= \quad \{w \in \mathbb{R}^{d_L d_0} : M^{\top}w = \mathbf{0}\}.
\end{aligned}$$

Following our convention in this paper for multiplying a matrix by a subspace, we can write the simpler definitions col $M = M\mathbb{R}^{d_\theta}$ and row $M = M^{\top}\mathbb{R}^{d_L d_0}$.

The differential map $d\mu(\theta)$ has the same four fundamental subspaces, except that its columnspace (image) and left nullspace are subspaces of $\mathbb{R}^{d_L \times d_0}$ instead of $\mathbb{R}^{d_L d_0}$. As $d\mu(\theta)$ is a linear map, it also has a transpose, which we write as $d\mu^{\top}(\theta)$. Considering that M represents the same linear transformation as $d\mu(\theta)$, let $d\mu^{\top}(\theta)$ be the linear map from $\mathbb{R}^{d_L \times d_0}$ to $\mathbb{R}^{d_\theta}$ represented by $M^{\top}$. Recall from (8.1) that, given a weight displacement $\Delta\theta = (\Delta W_L, \Delta W_{L-1}, \ldots, \Delta W_1)$,

$$d\mu(\theta)(\Delta\theta) = \sum_{j=1}^{L} W_{L\sim j} \Delta W_j W_{j-1\sim 0}.$$

Clearly this transformation is linear in $\Delta\theta$, but it isn't written as a multiplication by a single matrix M. Nevertheless, we can write its transpose, applied to a matrix ΔW, as

$$\mathrm{d}\mu^\top(\theta)(\Delta W) = (\Delta W W_{L-1\sim0}^\top, \ldots, W_{L\sim j}^\top \Delta W W_{j-1\sim0}^\top, \ldots, W_{L\sim1}^\top \Delta W).$$

The nullspace, columnspace (image), rowspace, and left nullspace of $\mathrm{d}\mu(\theta)$ are

$$
\begin{aligned}
\mathrm{null}\,\mathrm{d}\mu(\theta) &= \{\Delta\theta \in \mathbb{R}^{d_\theta} : \mathrm{d}\mu(\theta)(\Delta\theta) = 0\} = \mathrm{null}\,M, \\
\mathrm{col}\,\mathrm{d}\mu(\theta) = \mathrm{image}\,\mathrm{d}\mu(\theta) &= \{\mathrm{d}\mu(\theta)(\Delta\theta) : \Delta\theta \in \mathbb{R}^{d_\theta}\}, \\
\mathrm{row}\,\mathrm{d}\mu(\theta) = \mathrm{image}\,\mathrm{d}\mu^\top(\theta) &= \{\mathrm{d}\mu^\top(\theta)(\Delta W) : \Delta W \in \mathbb{R}^{d_L\times d_0}\} = \mathrm{row}\,M, \quad \text{and} \\
\mathrm{null}\,\mathrm{d}\mu^\top(\theta) &= \{\Delta W \in \mathbb{R}^{d_L\times d_0} : \mathrm{d}\mu^\top(\theta)(\Delta W) = \mathbf{0}\}.
\end{aligned}
$$

The only difference between $\mathrm{col}\,\mathrm{d}\mu(\theta)$ and $\mathrm{col}\,M$ is whether it is formatted as a set of $d_L \times d_0$ matrices or a set of vectors of length $d_L d_0$; and likewise for the difference between $\mathrm{null}\,\mathrm{d}\mu^\top(\theta)$ and $\mathrm{null}\,M^\top$.

We now apply some basic facts from linear algebra to the differential map.

- The fundamental subspaces $\mathrm{col}\,\mathrm{d}\mu(\theta)$ and $\mathrm{row}\,\mathrm{d}\mu(\theta)$ have the same dimension, called the *rank* of $\mathrm{d}\mu(\theta)$ and denoted $\mathrm{rk}\,\mathrm{d}\mu(\theta)$. (The rank can also be identified by the fact that, letting $r = \mathrm{rk}\,\mathrm{d}\mu(\theta)$, M has an $r \times r$ minor with a nonzero determinant but every $(r+1) \times (r+1)$ minor has determinant zero.)

- The rowspace of $\mathrm{d}\mu^\top(\theta)$ is the columnspace of $\mathrm{d}\mu(\theta)$ and the columnspace of $\mathrm{d}\mu^\top(\theta)$ is the rowspace of $\mathrm{d}\mu(\theta)$, so $\mathrm{rk}\,\mathrm{d}\mu^\top(\theta) = \mathrm{rk}\,\mathrm{d}\mu(\theta)$.

- $\mathrm{row}\,\mathrm{d}\mu(\theta)$ is the orthogonal complement of $\mathrm{null}\,\mathrm{d}\mu(\theta)$ (in the space $\mathbb{R}^{d_\theta}$). By the Rank-Nullity Theorem, $\mathrm{rk}\,\mathrm{d}\mu(\theta) + \dim\mathrm{null}\,\mathrm{d}\mu(\theta) = d_\theta$.

- $\mathrm{col}\,\mathrm{d}\mu(\theta)$ is the orthogonal complement of $\mathrm{null}\,\mathrm{d}\mu^\top(\theta)$ (in the space $\mathbb{R}^{d_L\times d_0}$; the orthogonality is in the Frobenius inner product). By the Rank-Nullity Theorem applied to $\mathrm{d}\mu^\top(\theta)$, $\mathrm{rk}\,\mathrm{d}\mu(\theta) + \dim\mathrm{null}\,\mathrm{d}\mu^\top(\theta) = d_L d_0$.

A major concern of this paper is to determine $\mathrm{null}\,\mathrm{d}\mu(\theta)$ (Section 8.5) and $\mathrm{row}\,\mathrm{d}\mu(\theta)$ (Section 9.1). Trager, Kohn, and Bruna [21] are more interested in $\mathrm{col}\,\mathrm{d}\mu(\theta)$, from which they derive $\mathrm{rk}\,\mathrm{d}\mu(\theta)$, which we use in Section 8.5 (Theorem 40) to show that our expression for $\mathrm{null}\,\mathrm{d}\mu(\theta)$ is correct—specifically, that we did not overlook any dimensions of $\mathrm{null}\,\mathrm{d}\mu(\theta)$. Hence we reprise their derivations of $\mathrm{col}\,\mathrm{d}\mu(\theta)$ and $\mathrm{rk}\,\mathrm{d}\mu(\theta)$ here (with some changes). Recall our conventions that $W_{j\sim j}$ is the $d_j \times d_j$ identity matrix, $\mathrm{col}\,W_{L\sim L} = \mathbb{R}^{d_L}$, $\mathrm{rk}\,W_{L\sim L} = d_L$, $\mathrm{row}\,W_{0\sim0} = \mathbb{R}^{d_0}$, and $\mathrm{rk}\,W_{0\sim0} = d_0$.

Lemma 57.

$$\mathrm{col}\,\mathrm{d}\mu(\theta) = \sum_{j=1}^{L} \mathrm{col}\,W_{L\sim j} \otimes \mathrm{row}\,W_{j-1\sim0} \tag{C.1}$$

Proof. Recall again that $\mathrm{d}\mu(\theta)(\Delta\theta) = \sum_{j=1}^{L} W_{L\sim j}\Delta W_j W_{j-1\sim0}$. Let $(\Delta W_j)_{st}$ denote the component in row s and column t of ΔW_j, let C_s be column s of $W_{L\sim j}$ (expressed as a column vector), and let R_t be row t of $W_{j-1\sim0}$ (expressed as a row vector). Then $W_{L\sim j}\Delta W_j W_{j-1\sim0} = \sum_s \sum_t (\Delta W_j)_{st} C_s R_t$, which pairs each column of $W_{L\sim j}$ with each row of $W_{j-1\sim0}$ to form an outer product matrix, with an independent coefficient for each pairing. It follows that the term $W_{L\sim j}\Delta W_j W_{j-1\sim0}$ can independently take on any value in $\mathrm{col}\,W_{L\sim j} \otimes \mathrm{row}\,W_{j-1\sim0}$, so the image of $\mathrm{d}\mu(\theta)$ is $\sum_{j=1}^{L} \mathrm{col}\,W_{L\sim j} \otimes \mathrm{row}\,W_{j-1\sim0}$. $\qquad\square$

Lemma 58 (Trager, Kohn, and Bruna [21], Lemma 3).

$$\text{rk } d\mu(\theta) = \sum_{j=1}^{L} \text{rk } W_{L\sim j} \cdot \text{rk } W_{j-1\sim 0} - \sum_{j=1}^{L-1} \text{rk } W_{L\sim j} \cdot \text{rk } W_{j\sim 0}$$

Proof. The dimension of each summand in (C.1) is $\text{rk } W_{L\sim j} \cdot \text{rk } W_{j-1\sim 0}$, but the summands overlap. For any two subspaces X and Y, $\dim(X+Y) = \dim X + \dim Y - \dim(X \cap Y)$. When the summand $\text{col } W_{L\sim j} \otimes \text{row } W_{j-1\sim 0}$ is added, its intersection with the vector sum of the previous summands is $\text{col } W_{L\sim j-1} \otimes \text{row } W_{j-1\sim 0}$, so by induction the dimension of $\text{col } d\mu(\theta)$ is

$$\begin{aligned}
\text{rk } d\mu(\theta) &= \text{rk } W_{L\sim 1} \cdot \text{rk } W_{0\sim 0} + \sum_{j=2}^{L} \left(\text{rk } W_{L\sim j} \cdot \text{rk } W_{j-1\sim 0} - \text{rk } W_{L\sim j-1} \cdot \text{rk } W_{j-1\sim 0} \right) \\
&= \sum_{j=1}^{L} \text{rk } W_{L\sim j} \cdot \text{rk } W_{j-1\sim 0} - \sum_{j=1}^{L-1} \text{rk } W_{L\sim j} \cdot \text{rk } W_{j\sim 0}.
\end{aligned}$$

$\square$

Corollary 59. *The dimension of* $\text{null } d\mu(\theta)$ *is* D^{free}, *as specified by (8.19)*

Proof. By the Rank-Nullity Theorem, $\dim \text{null } d\mu(\theta) = d_\theta - \text{rk } d\mu(\theta) = D^{\text{free}}$. $\square$

D   Counting More Dimensions

Table 7 defines several sets of subspaces that were not important enough to include in Section 8.7, and the dimensions of the subspaces (of $\mathbb{R}^{d_\theta}$) they span.

Let's count the degrees of freedom of connecting moves. Recall from Section 7.4 that the set of subspaces associated with connecting moves is $\Theta_{\text{O}}^{\text{conn}} = \{\phi_{l,j-1,j,j,h} \neq \{\mathbf{0}\} : L \geq l \geq j > h \geq 0\}$. First we count the degrees of freedom associated with connecting moves that change one particular matrix W_j:

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

The space spanned by $\Theta_{\text{O}}^{\text{conn}}$ has dimension $D_{\text{O}}^{\text{conn}} = \sum_{j=1}^{L} D_j^{\text{conn}}$. Thus we obtain the formula (D.1) (see Table 7).

Let's count the degrees of freedom of swapping moves. Recall from Section 7.4 that the set of subspaces associated with swapping moves is $\Theta_{\text{O}}^{\text{swap}} = \{\phi_{lkjih} \in \Theta_{\text{O}} : l > k \geq i > h\}$. The dimension $D_{\text{O}}^{\text{swap}}$ of the space spanned by $\Theta_{\text{O}}^{\text{swap}}$ can be derived exactly as we derived $D_{\text{O}}^{\text{comb}}$ in Section 8.7, except that we omit the subspaces where $k = i - 1$, which indicate connecting moves and not swapping moves. Thus we obtain the formula (D.2) (see Table 7).

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 $\Theta_{\text{O}}^{\text{fiber,comb}} = \Theta_{\text{O}}^{\text{fiber}} \cap \Theta_{\text{O}}^{\text{comb}}$. The easiest way to determine the dimension of the subspace spanned by $\Theta_{\text{O}}^{\text{fiber,comb}}$ is to first understand the prebasis $\Theta_{\text{O}}^{L0,\text{comb}} = \Theta_{\text{O}}^{L0} \cap \Theta_{\text{O}}^{\text{comb}}$, which is the set

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

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

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

$$D_O^{\text{swap}} = \sum_{L > k \geq i > 0} (k - i + 2)\,\underbrace{(\text{rk } W_{k+1\sim i} - \text{rk } W_{k+1\sim i-1})}_{\beta_{k+1,i,i}}\,\underbrace{(\text{rk } W_{k\sim i-1} - \text{rk } W_{k+1\sim i-1})}_{\alpha_{k,k,i-1}} \tag{D.2}$$

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

$$D_O^{L0,\text{comb}} = \sum_{L \geq k+1 \geq i > 0} (k - i + 2)\,\omega_{Li}\,\omega_{k0} \tag{D.3}$$

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

$$D_O^{\text{fiber,comb}} = D_O^{\text{comb}} - D_O^{L0,\text{comb}} = \sum_{L \geq k+1 \geq i > 0} (k - i + 2)\,(\beta_{k+1,i,i}\,\alpha_{k,k,i-1} - \omega_{Li}\,\omega_{k0}) \tag{D.4}$$

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

$$D_O^{L0,\text{conn}} = \sum_{j=1}^{L} \omega_{Lj}\,\omega_{j-1,0} = \sum_{j=1}^{L} \underbrace{(\text{rk } W_{L\sim j} - \text{rk } W_{L\sim j-1})}_{\omega_{Lj}}\,\underbrace{(\text{rk } W_{j-1\sim 0} - \text{rk } W_{j\sim 0})}_{\omega_{j-1,0}} \tag{D.5}$$

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

$$D_O^{\text{fiber,conn}} = \sum_{j=1}^{L} \left(\underbrace{(d_j - \text{rk } W_j)}_{\beta_{jjj}}\,\underbrace{(d_{j-1} - \text{rk } W_j)}_{\alpha_{j-1,j-1,j-1}} - \underbrace{(\text{rk } W_{L\sim j} - \text{rk } W_{L\sim j-1})}_{\omega_{Lj}}\,\underbrace{(\text{rk } W_{j-1\sim 0} - \text{rk } W_{j\sim 0})}_{\omega_{j-1,0}} \right) \tag{D.6}$$

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

$$D_T^{L0,\text{comb}} = \sum_{L > k \geq i > 0} (k - i + 1)\,\omega_{Li}\,\omega_{k0} \tag{D.7}$$

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

$$D_T^{\text{fiber,comb}} = D_T^{\text{comb}} - D_T^{L0,\text{comb}} = \sum_{L > k \geq i > 0} (k - i + 1)\,(\beta_{k+1,i,i}\,\alpha_{k,k,i-1} - \omega_{Li}\,\omega_{k0}) \tag{D.8}$$

Table 7: More sets of subspaces of $\mathbb{R}^{d_\theta}$ and their total dimensions. See also Table 6.

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

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

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

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

The dimension of the space spanned by $\Theta_O^{\text{fiber,comb}}$ is

$$
\begin{aligned}
D_O^{\text{fiber,comb}} &= \dim \Theta_O^{\text{fiber,comb}} \\
&= \dim \left(\Theta_O^{\text{comb}} \setminus \Theta_O^{L0} \right) \\
&= \dim \left(\Theta_O^{\text{comb}} \setminus \Theta_O^{L0,\text{comb}} \right) \\
&= D_O^{\text{comb}} - D_O^{L0,\text{comb}},
\end{aligned}
$$

from which we obtain the formula (D.4) in Table 7. Analogously, we define $\Theta_T^{\text{fiber,comb}} = \Theta_T^{\text{comb}} \setminus \Theta_T^{L0}$, which spans a space of dimension $D_T^{\text{fiber,comb}} = D_T^{\text{comb}} - D_T^{L0,\text{comb}}$, from which we obtain the formula (D.8).

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

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

The dimension of the space spanned by $\Theta_O^{L0,\text{conn}} = \Theta_O^{L0} \cap \Theta_O^{\text{conn}}$ is

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

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

$$
\begin{aligned}
D_O^{\text{fiber,conn}} &= D_O^{\text{conn}} - D_O^{L0,\text{conn}} \\
&= \sum_{j=1}^{L} \left((d_j - \text{rk } W_j)\,(d_{j-1} - \text{rk } W_j) - (\text{rk } W_{L\sim j} - \text{rk } W_{L\sim j-1})\,(\text{rk } W_{j-1\sim 0} - \text{rk } W_{j\sim 0}) \right).
\end{aligned}
$$

Perhaps it is worth noting that the triple summation in (8.11) 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 (8.16), the term $\omega_{Li}\,\omega_{k0}$ occurs once for each $j \in [\max\{i,1\}, \min\{k, L-1\}]$. Hence, we can write

$$
D_O^{L0} = \sum_{i=0}^{L} \sum_{k=\max\{i-1,0\}}^{L} (\min\{k+1, L\} - \max\{i-1, 0\})\,\omega_{Li}\,\omega_{k0} \quad \text{and}
$$

$$
D_T^{L0} = \sum_{i=0}^{L-1} \sum_{k=\max\{i,1\}}^{L} (\min\{k, L-1\} - \max\{i-1, 0\})\,\omega_{Li}\,\omega_{k0}.
$$

E An Affine Transformation of the Nullspace of the Differential Map

Consider a point θ on a fiber $\mu^{-1}(W)$, and recall from Section 5.2 the linear function $\eta : (M_L, M_{L-1}, \ldots, M_1) \mapsto (J_L M_L J_{L-1}^{-1}, J_{L-1} M_{L-1} J_{L-2}^{-1}, \ldots, J_1 M_1 J_0^{-1})$ which maps θ to a canonical weight vector $\tilde\theta$ (with the same rank list). (The J matrices depend on θ.) Recall Lemma 13, which shows that $\eta(\mu^{-1}(\tilde I)) = \mu^{-1}(W))$. We now show that η also maps the nullspace of $d\mu(\tilde\theta)$ to the nullspace of $d\mu(\theta)$. This is not very surprising, as invertible affine transformations preserve tangencies; but it's nice to have a formal proof.

Lemma 60. $\eta\left(\text{null d}\mu(\tilde{\theta})\right) = \text{null d}\mu(\theta)$.

Proof. Recall the formula (8.1) for $d\mu(\theta)$. Substituting in $\tilde{\theta}$ gives

$$d\mu(\tilde{\theta})(\Delta\theta) = \sum_{j=1}^{L} \tilde{I}_{L\sim j} \Delta W_j \tilde{I}_{j-1\sim 0} = J_L^{-1} \left(\sum_{j=1}^{L} W_{L\sim j} J_j \Delta W_j J_{j-1}^{-1} W_{j-1\sim 0} \right) J_0 = J_L^{-1} d\mu(\theta)(\eta(\Delta\theta)) J_0.$$

For any $\Delta\theta \in \mathbb{R}^{d_\theta}$, $\Delta\theta \in \text{null d}\mu(\tilde{\theta})$ means that $d\mu(\tilde{\theta})(\Delta\theta) = 0$, which is true if and only if $d\mu(\theta)(\eta(\Delta\theta)) = 0$, or equivalently, $\eta(\Delta\theta) \in \text{null d}\mu(\theta)$. Therefore, $\eta\left(\text{null d}\mu(\tilde{\theta})\right) = \text{null d}\mu(\theta)$ as claimed. □

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] Riccardo Benedetti and Jean-Jacques Risler. *Real Algebraic and Semi-Algebraic Sets*. Hermann, Éditeurs des Sciences et des Arts, Paris, 1990.

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

[5] 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.

[6] 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.

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

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

[9] Joe Harris. *Algebraic Geometry: A First Course*, Graduate Texts in Mathematics, volume 133. Graduate Texts in Mathematics. Springer, New York, 1995.

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

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

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

[13] Stanisław Łojasiewicz. *Ensembles Semi-Analytiques*. Unpublished manuscript. A version typeset by Michel Coste is available at https://perso.univ-rennes1.fr/michel.coste/Lojasiewicz.pdf, July 1970.

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

[15] Yung-Chen Lu. *Singularity Theory and an Introduction to Catastrophe Theory*. Springer Verlag, New York, 1976.

[16] John Mather. *Notes on Topological Stability*. Lecture Notes. Reprinted in Bulletin of the American Mathematical Society **49**(4):475–506, October 2012, 1970.

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

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

[19] René Thom. *La Stabilité Topologique des Applications Polynomiales*. L'Enseignement Mathématique **8**(1–2):24–33, 1962.

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

[21] 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.

[22] Hassler Whitney. *Elementary Structure of Real Analytic Varieties*. Annals of Mathematics **66**(2):545–556, 1957.

[23] ———. *Local Properties of Analytic Varieties*. Differential and Combinatorial Topology (Stewart Scott Cairns, editor), pages 205–244. Princeton University Press, 1965.

[24] ———. *Tangents to an Analytic Variety*. Annals of Mathematics **81**(2):496–549, 1965.

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