Pure and Applied Mathematics Quarterly
Volume 19, Number 2, 731–789, 2023

Noncommutative geometry of computational models and uniformization for framed quiver varieties

George Jeffreys and Siu-Cheong Lau

Abstract: We formulate a mathematical setup for computational neural networks using noncommutative algebras and near-rings, in motivation of quantum automata. We study the moduli space of the corresponding framed quiver representations, and find moduli of Euclidean and non-compact types in light of uniformization.

Keywords: Noncommutative geometry, near-rings, neural networks, deep learning, representation theory, moduli spaces.

1. Introduction

The connections between computer science and algebra are profound. In the early 1900s, both were deeply tied to practical and philosophical developments towards understanding what it truly means to calculate something. For example, there was Turing's Halting problem and Gödel's Incompleteness theorem.

As modern abstract algebra was developed in the 50s and 60s, it was fruitfully applied towards computer science with the creation of the theory of finite automata. The first fundamental result in this development was Kleene's Theorem demonstrating that the class of recognizable languages is the class of rational languages [21]. In 1956, Schützenberger defined the *syntactic monoid*, a canonical monoid attached to each language [29]. Later, he proved that a language is star-free exactly when its syntactic monoid is finite and aperiodic [30]. At this point mathematicians started to consider the algebraic geometry of these monoids as Birkhoff [4] and later Eilenberg [14] and Reiterman [27] wrote about varieties of these monoids (infinite and finite respectively).

The theory of finite automata arose from an extremely widespread interdisciplinary effort to understand calculation. Modern science suggests that the brain operates as a so-called neural network, the structure of which has inspired the computational tool known as the artificial neural network. Neural network models heavily use graphs and their linear representations. This gives rise to further relations between mathematics and computer science.

Received May 30, 2022.

In this paper, enlightened by the modern developments of quantum mechanics and the applications in computer science, we would like to further study relationships between these subjects by constructing an algebraic geometric model for both neural networks and quantum automata. We hope that this framework can partially reveal deep relations between these two seemingly distant areas. Our study is just the tip of the iceberg of connections between these subjects. There are various deep topics that we have not touched yet, such as analysis of the large N limit by taking the representing dimensions to infinity, stochastic analysis for the geometry of quiver moduli, dynamical systems in relation with recurrent neural networks, and so on.

Summary of results

This paper is theoretical in nature. The main outcome is a mathematical framework using quiver near-algebras and metrics over moduli that can formulate both quantum automata and deep learning algorithms. Definition 1.1 gives such an algebraic model. In particular, it provides a physical interpretation of operations in a network model in terms of quantum measurement.

We make a systematic study of representations of near-rings and construct quiver near-rings. We construct differential forms over a near-ring and show that they induce differential forms over moduli of representations in all dimensions with values in $\mathrm{Map}(F, F)$, the space of maps on the framing (Theorem 1.2). Zero-forms and one-forms are the basic building blocks. In a deep learning model, they are used to encode the cost function and its differential.

Moreover, we construct an interpolation between metrics on the compact framed moduli and the Euclidean space. As a result, the usual Euclidean formulation of deep learning is included as a special instance in our framework. We also constructed framed quiver moduli of hyperbolic type in Theorem 1.3. In Section 4.5, we provide explicit formulas and simplifications for implementing the model. There are many recent studies that investigate computational efficiency and advantages for non-Euclidean learning. Our paper is mainly for theoretical purpose and does not include experiments that test for efficiency.

More detailed descriptions

A finite automata consists of a set of states of a machine, a set of transitions between the states, and an alphabet set that will form a machine language, whose elements label the transitions of states. A quantum version of this replaces the set of states by a collection of vector spaces whose elements are known as state vectors. The set of transitions is replaced by a set of linear

maps between the vector spaces. This forms a so-called quiver representation, which is a linear representation of the directed graph Q (called a quiver) whose vertices label the collection of vector spaces, and whose arrows label the set of linear maps.

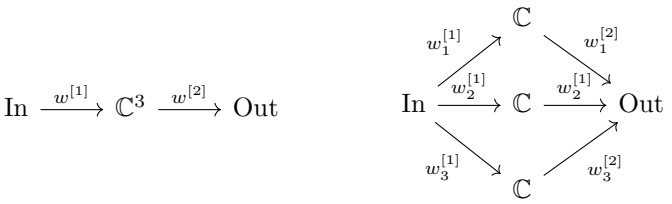

Figure 1: Two artificial neural networks with similar graphs. The LHS is known as an A_3-quiver.

Paths in the quiver play the role of words of a machine language. The path algebra

$$\mathcal{A} = \mathbb{C}Q$$

consists of complex linear combinations of paths, with concatenation of paths serving as the product. Taking linear combinations can be interpreted as forming superpositions of quantum states.

In summary, *a quiver algebra and its modules provide an algebraic model of a quantum automata*.

One crucial component that one cannot miss is *taking observation of the quantum particles*. Most mathematical physics literature concentrates on the quantum propagation process, and have left away the mysterious observation step, perhaps due to its probabilistic and singular nature. However, this step is crucial in true understanding of quantum physics, and also in practical applications. For modeling quantum propagations, operator algebras serve as a very successful mathematical tool. However, to include the observation process, we find that a *near-ring*, which is much less studied than an algebra, is necessary.

To model the observation process in a quantum world, we need two more ingredients: a Hermitian metric h of the state space V, and a framing linear map $e : F \to V$ where $F = \mathbb{C}^n$ is called a framing vector space. Then we take

$$e^{*_h}(v) = \sum_{j=1}^{n} h(e(\epsilon_j), v)\epsilon_j^*,$$

where ϵ_j denotes the standard basis of $\mathbb{C}^n$ (and ϵ_j^* denotes the dual basis). The coefficients $h(e(\epsilon_j), v)$ are interpreted as the quantum amplitudes of a state

v being $e(\epsilon_j)$. Then the quantum collapsing after observation is modeled by composing this with a fixed non-linear activation function $\sigma : F \to F$ (for instance a certain step function, or a smoothing of it). In the quantum world, σ is indeed an F-valued probability distribution on F.

Thus, a quantum machine consists of not just linear transitions of states, but also the framings and non-linear activation functions that correspond to taking observations. We will make the following definition. See also Figure 3.

Definition 1.1 (Definition 3.5)**.** *An* activation module *consists of:*

1. *a (noncommutative) algebra $\mathcal{A}$ and vector spaces V, $F = F_{\text{in}} \oplus F_{\text{out}} \oplus F_{\text{m}}$; ('m' stands for 'memory' or 'middle'.)*
2. *A family of metrics $h_{(w,e)}$ on V over the space of framed $\mathcal{A}$-modules*

$$R = \text{Hom}_{\text{alg}}(\mathcal{A}, \text{End}(V)) \times \text{Hom}(F, V)$$

   which is $\text{GL}(V)$-equivariant;
3. *a collection of possibly non-linear functions*

$$\sigma_j^F : F_{\text{m}} \to F_{\text{m}}.$$

In above, R parametrizes computing machines that have the same underlying framed quiver, and hence is governed by the same language. Moreover, framed $\mathcal{A}$-modules that differ by a $\text{GL}(V)$-action have the same computational effect and hence should be identified. $[R/\text{GL}(V)]$ forms a moduli stack of computing machines.

In this formulation, a machine language is composed of not just linear transitions of state spaces, but also non-linear (or probabilistic) operations σ that models quantum observations. The set of operations generated by these is no longer an algebra, since

$$\sigma \circ (\gamma_1 + \gamma_2) \neq \sigma \circ \gamma_1 + \sigma \circ \gamma_2$$

where γ_1, γ_2 are composed of linear operations in $\mathcal{A}$ and the dual framing map e^{*_h}. Rather, it generates a near-algebra $\widetilde{\mathcal{A}}$, which is almost a ring except that the multiplication (which is realized by composition of maps in the current setup) fails to be distributive on one side.

In our formulation, representations of $\widetilde{\mathcal{A}}$ play a key role in computational models. However, the space of representations of a near-ring was not studied in previous literature. In this paper, we begin to investigate some aspects of the representation theory for a near-ring. We construct quiver near-algebras which are new mathematical objects to the authors' knowledge. From the

standpoint of machine learning, the near-algebra universally controls all the machines with the same underlying network in all dimensions at the same time. We are not claiming that the construction will directly lead to more effective algorithms. Rather, it aims to provide a universal algebraic model for both quantum automata and deep learning.

In noncommutative geometry, a noncommutative ring is understood as a 'space', in analogous to $\mathrm{Spec}(A)$ for a commutative ring A. For a noncommutative ring A, differential forms and their cohomology theory found by Connes [11], Cuntz-Quillen [12] and Ginzburg [18] classify the deformations of A. We would like to extend the construction to the context of near-rings.

The main idea is that, every element in the near-ring $\widetilde{\mathcal{A}}$, which is interpreted as a program written in the language of $\widetilde{\mathcal{A}}$, produces a family of maps on the framing space F over the moduli of machines $[R/G]$. In other words, each machine parametrized by a point in $[R/G]$ performs a computation $F \to F$ specified by the program. We construct differential forms for a near-ring $\widetilde{\mathcal{A}}$ and extend this association from $\widetilde{\mathcal{A}}$ to its space of representations $[R/G]$.

Theorem 1.2 (Theorem 3.42)**.** *There exists a degree-preserving map*

$$DR^{\bullet}(\widetilde{\mathcal{A}}) \to (\Omega^{\bullet}(R, \mathbf{Map}\,(F, F)))^{G}$$

which commutes with d on the two sides. In above, $\mathbf{Map}\,(F, F)$ *denotes the trivial bundle* $\mathrm{Map}\,(F, F) \times R$ *where* $\mathrm{Map}\,(F, F)$ *is the set of* $\mathbb{C}$*-valued smooth maps from* F *to itself, and the action of* $G = \mathrm{GL}(V)$ *on fiber direction is trivial.*

The main difference from the case of rings is that the near-ring $\widetilde{\mathcal{A}}$ is framed, and it contains non-linear elements σ at the framing. As a consequence, the right hand side of above takes values in the (infinite-dimensional) space of maps on the framing.

For general differential forms, 0-forms and 1-forms are the basic building blocks. In machine learning, for a fixed algorithm $\tilde{\gamma} \in \widetilde{\mathcal{A}}$, a learning process attempts to find a machine $p \in [R/G]$ that produces the best fit computation $\phi_p^{\tilde{\gamma}} : F \to F$ by minimizing a certain 0-form (for instance $\int_K \left|\varphi_p^{\tilde{\gamma}}(x) - f\,(x)\right|^2 dx$ for a given $f : K \to \mathbb{R}$ and $K \subset F$ in supervised learning). Its differential, which is a 1-form in $DR^1(\widetilde{\mathcal{A}})$, governs the gradient flow on $[R/G]$ with the help of a metric.

In general, $[R/G]$ is a singular stack. Fortunately, when the dimension vector is primitive, one can construct a fine moduli of framed quiver representations by taking a GIT quotient (with respect to a suitably chosen

stability condition) [20, 24]. Such moduli spaces $\mathcal{M}$ can be used in place of $[R/G]$. When the quiver has no oriented cycle (called to be acyclic), the moduli $\mathcal{M}$ is compact. The topology of framed quiver moduli is well studied by [26].

In [19], we formulated learning by neural networks over a framed moduli space $\mathcal{M}$. Namely, the state space V_i over each vertex $i \in Q_0$ patches up as a universal bundle $\mathcal{V}_i$ over $\mathcal{M}$. The transition arrows $a \in Q_1$ correspond to bundle maps over $\mathcal{M}$. The framing linear maps $e_i : F_i \to V_i$ correspond to bundle maps from the trivial bundle $\mathbf{F_i}$ to $\mathcal{V}_i$. Then data and states of the family of machines are naturally modeled by sections over $\mathcal{M}$; propagation of signals is modeled by bundle maps. In this formulation, learning is a stochastic gradient descent over the moduli $\mathcal{M}$.

It is tempting to ask how this formulation relates to the most common method of machine learning over an Euclidean space, rather than a moduli space $\mathcal{M}$. In this paper, we will answer this question in light of uniformization of metrics.

The moduli space $\mathcal{M}$ is topologically a compactification of the Euclidean space, now denoted as $\mathcal{M}^0$. The main observation is that, the Euclidean space $\mathcal{M}^0$ can be interpreted as a moduli space of positive-definite quiver representations with respect to a certain Hermitian form H_i^0 for the universal bundles $\mathcal{V}_i$. Thus, the most popular approach using Euclidean space indeed also falls into our formulation of learning over the moduli space.

This uniformization picture naturally includes a hyperbolic version of the moduli space. Namely, by changing the signature of the quadratic form (see (23)), we obtain another type of moduli space $\mathcal{M}^-$ of positive-definite quiver representations with respect to H_i^-. We show that $\mathcal{M}^-$ comes with a natural metric.

Theorem 1.3 (Theorem 4.15)**.** *Define H_T^- to be $H_T^- := -i\sum_i \partial\overline{\partial}\log\det H_i^-$ on $\mathcal{M}^-$. Then H_T^- is a Kähler metric on $\mathcal{M}^-$.*

In typical applications, one usually restrict to real coefficients. Correspondingly, the formulae provided by this paper give bundle metrics for $\mathcal{V}_i^{\mathbb{R}}|_{\mathcal{M}_{\mathbb{R}}}$ and Riemannian metrics on $\mathcal{M}_{\mathbb{R}}$.

As a result, we can run machine learning over $\mathcal{M}, \mathcal{M}^0, \mathcal{M}^-$, or an interpolation of them. We can also set learnable parameters that interpolate these spaces, and let the machine learn which metric serves the best for a given task. We discuss more in Section 4.5 for concrete implementation of these moduli spaces in deep learning algorithms.

Some related works

Recently, there is a rising interest in the connections between neural networks and quiver representations. The paper [2] found a new way of encoding the data flow as a quiver representation, which makes a crucial use of the assumption of thin representations (where dimensions of representing vector spaces over vertices are all 1). This was extended in [1] which used moduli spaces of doubly framed quiver representations. On the other hand, the gradient descent that they take is not directly carried out over the quiver moduli, and hence is different from our approach in [19] and this paper. The paper [17] studied the symmetries coming from the quiver approach to neural networks.

There are also newly invented approaches to apply modern mathematics to machine learning. Most literature concerns about the input data set and endows it with additional mathematical structures, for instance, Lie group symmetry [10, 8, 7, 9, 6, 13], or categorical structures [31]. On the other hand, in our current approach, we focus on the computing machine itself, and formulate its algebro-geometric structure and makes use of its internal symmetry.

For learning using hyperbolic spaces, there are several beautiful works, see for instance [25], [15], [28], [16]. The non-compact dual of the moduli space $\mathcal{M}^-$ that we introduce in this paper can be understood as a higher rank generalization of hyperbolic spaces in the sense of Hermitian symmetric spaces. See more in Section 4.4.

Organization of this paper

In Section 2, we present the motivating example for the machine learning applications in this paper. In Section 3, we will define computing machines in the context of noncommutative geometry. In Section 4, we will apply the idea of uniformization of metrics to construct non-compact duals to neural network quiver moduli spaces.

2. A guiding example

Let's first review the basic setup of deep learning via the following typical real-life example. The MNIST dataset is a collection of 70,000 handwritten single digit numbers, each one stored as a 28-pixel-by-28-pixel image, along with a label in $\{0, \ldots, 9\}$, namely the digit shown by the picture. It is a common first experiment in machine learning to build a simple neural network and to use the MNIST dataset to train it to recognize handwritten digits.

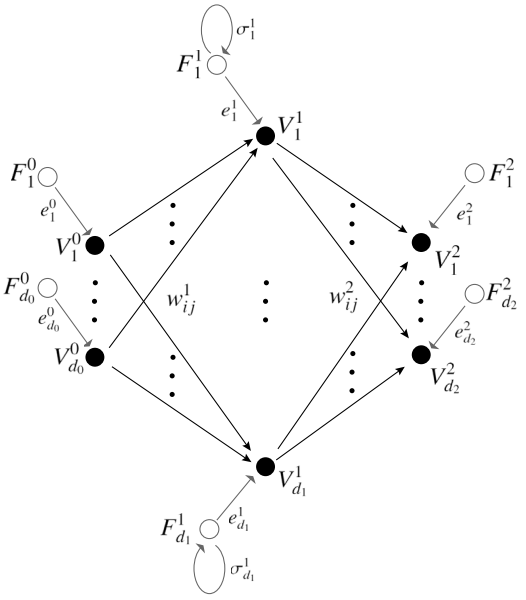

Figure 2

Typically, the network will have three layers as in Figure 2. The first one is the input layer which has $d_0 = 784$ neurons, one for each pixel in the 28 x 28 image. The input for each neuron is the intensity of the corresponding pixel (ranging from 0 (white) to 255 (black)). The second layer is the so called "hidden layer," the layer where the non-trivial calculations happen. Let d_1 denote the number of neurons in the hidden layer. The third layer is the output layer, which has $d_2 = 10$ neurons corresponding to digits 0 to 9. The network is trained to recognize an image of a handwritten digit and output the vector corresponding to that digit. For example, an image of '5' should result in the column vector $[0, 0, 0, 0, 0, 1, 0, 0, 0, 0]^T$.

We will use the vector space V_i^k (currently one-dimensional) to represent the i-th neuron in layer k, where $k = 0$ is the input layer, $k = 1$ is the hidden layer, and $k = 2$ is the output layer. The weight from the $(k-1, j)$-neuron to neuron (k, i) will be denoted w_{ij}^k. The activation function and bias applied to $(1, i)$-neuron will be denoted σ_i^1 and b_i respectively. Typically, σ_i^1 is the same for all i, and is denoted by σ in this case. (For the moment, we ignore F_i^k (called the framing) and put the activation functions directly on the state spaces V_i^k. Later on, we will explain why we introduce framing and put the activation functions on F_i^k instead.)

The image is sent to the input layer as a vector z^0 with $d_0 = 784$ entries, each corresponding to a neuron in the input layer. The signals in the input layer are sent to the neurons in the hidden layer via the affine linear map $w^1 z^0 + b$, giving the pre-activation vector z^1. Here, $w^1 = (w^1_{ij})$ is the $d_1 \times 784$ matrix and $b = (b_i)$ is the $d_1 \times 1$ vector of biases. The activation functions convert the pre-activation vector z^1 to the activation vector a^1 via $a^1_i = \sigma(z^1_i)$ applied component-wise. Finally, the vector a^1 is sent to the output neurons by the linear map w^2, which is a $10 \times d_1$ matrix. This results in the output vector z^2.

At the start, the weight matrices w^1, w^2 and the bias vector b are initialized as some random values. The goal is to optimize these parameters. This is done with respect to an overall cost function which measures the distance between the output vector and the correct vector for all sample images. Let L be the domain of images. The cost function can be taken to be

$$\mathcal{C}(w^1, w^2, b) = \int_{x \in L} \| f_{w^1, w^2, b}(x) - y_x \|^2$$

where $f_{(w^1, w^2, b)}$ is the function produced by the network with parameters w^1, w^2, b, and y_x is the vector associated to the labeled digit for input x.

In practice, a stochastic gradient descent is performed where these quantities are calculated and numerically optimized for some finite sample subset K of L. This is done via *forward propagation*, which consists of evaluating $f_{(w_1, w_2, b)}(x)$ at each $x \in K$ and *backpropagation*, which consists of using these values as well as the pre-determined derivative of σ to compute the relevant derivatives.

3. An algebro-geometric formulation of computing machine

In this section, we give a mathematical formulation of a computing machine based on algebra and geometry. First, we formulate a machine as a framed module over an algebra, together with a metric on the module and a collection of non-linear functions. Second, we take into account of isomorphisms of framed modules and make sure the construction is equivariant under the automorphism group, and hence descends to the moduli stack of framed modules. Finally, we extend the noncommutative geometry developed by [11, 12, 18] to the context of near-rings, and show how it fits into this framework.

Our formulation uses framed modules. It has the theoretical advantage of separating the basis-free state space, which receives linear actions that are crucial in neural networks or quantum computations, from the framing vector space, which is necessary for the action of non-linear activation functions or quantum projections.

3.1. Intuitive construction

In this section, we make the basic algebraic setup and recall some definitions.

Let $\mathcal{A}$ be an associative algebra over $\mathbb{C}$ with unit $1_{\mathcal{A}}$. We use the algebra to encode all possible linear operations of the machine. Later, in the context of neural networks, we will take $\mathcal{A}$ to be the path algebra of a directed graph (which is also called a quiver).

Let V be a finite-dimensional vector space. V is understood as the space of abstract states of the machine prior to any physical observation. It is basis-free, namely, we do not pick any preferred choice of basis.

We consider $\mathcal{A}$-module structures on V, which are algebra homomorphisms $w\colon \mathcal{A} \to \mathfrak{gl}(V)$ (where $\mathfrak{gl}(V)$ denotes the algebra of all endomorphisms of V). Each module structure w realizes $a \in \mathcal{A}$ as a linear operation on the state space.

In reality, data are observed and recorded in fixed basis. For this, we define a framing vector space $F = F_{\mathrm{in}} \oplus F_{\mathrm{out}} \oplus F_{\mathrm{m}}$. Each component is a vector space with a fixed basis. The spaces F_{in} and F_{out} are respectively the vector spaces of all possible inputs and outputs. The space F_{m} can be understood as a space for memory of the machine. We may simply write $F = \mathbb{C}^n$ with the standard basis. The dimensions of $F_{\mathrm{in}}, F_{\mathrm{out}}, F_{\mathrm{m}}$ are denoted by $n_{\mathrm{in}}, n_{\mathrm{out}}, n_{\mathrm{m}}$ respectively. Moreover, we consider linear maps $e\colon F \to V$, $e = e_{\mathrm{in}} + e_{\mathrm{out}} + e_{\mathrm{m}}$ which are called the framing maps. The framing maps e are used to observe and record the abstract states.

A triple (V, w, e) is called a framed $\mathcal{A}$-module. We denote by

$$R \coloneqq \{(w, e) : w\colon \mathcal{A} \to \mathrm{End}\,(V)\,;\ e\colon F \to V\}$$

the *set of framed modules*. It serves as the parameter space of the machine. R is a subvariety in $\mathrm{Hom}(\mathcal{A}, \mathrm{End}(V)) \times \mathrm{Hom}(F, V)$.

As explained above, $\mathcal{A}$ encodes the internal linear operations on the state space V of the machine. In order to include the operations of exchanging data between the internal space and memory, we enlarge $\mathcal{A}$ to $\mathcal{A}_{\mathrm{m}}$, where $\mathcal{A}_{\mathrm{m}}$ is the augmented algebra

$$\mathcal{A}_{\mathrm{m}} = \mathcal{A}\langle 1_{\mathrm{m}}, \mathfrak{e}_{\mathrm{m}}, \mathfrak{e}_{\mathrm{m}}^*\rangle / I \tag{1}$$

where I is the two-sided ideal generated by the relations

$$1_{\mathrm{m}} \cdot \mathfrak{e}_{\mathrm{m}},\ \mathfrak{e}_{\mathrm{m}} \cdot 1_{\mathrm{m}} - \mathfrak{e}_{\mathrm{m}},\ 1_{\mathcal{A}} \cdot \mathfrak{e}_{\mathrm{m}} - \mathfrak{e}_{\mathrm{m}},$$
$$\mathfrak{e}_{\mathrm{m}}^* \cdot 1_{\mathrm{m}},\ 1_{\mathrm{m}} \cdot \mathfrak{e}_{\mathrm{m}}^* - \mathfrak{e}_{\mathrm{m}}^*,\ \mathfrak{e}_{\mathrm{m}}^* \cdot 1_{\mathcal{A}} - \mathfrak{e}_{\mathrm{m}}^*,$$

$$\mathfrak{e}_{\mathrm{m}}^2,\ (\mathfrak{e}_{\mathrm{m}}^*)^2,\ a\cdot\mathfrak{e}_{\mathrm{m}}^*,\ \mathfrak{e}_{\mathrm{m}}\cdot a,\ a\cdot 1_{\mathrm{m}},\ 1_{\mathrm{m}}\cdot a$$

for all $a\in\mathcal{A}$. (This means, for instance, $1_{\mathrm{m}}\cdot\mathfrak{e}_{\mathrm{m}}=0$ and $\mathfrak{e}_{\mathrm{m}}\cdot 1_{\mathrm{m}}=\mathfrak{e}_{\mathrm{m}}$ in the algebra $\mathcal{A}_{\mathrm{m}}$.) The unit of $\mathcal{A}_{\mathrm{m}}$ is $1_{\mathcal{A}}+1_{\mathrm{m}}$. $\mathfrak{e}_{\mathrm{m}}$ will be realized as a map from memory to the state space; $\mathfrak{e}_{\mathrm{m}}^*$ models saving results from the state space to memory.

Let's equip V with a Hermitian metric h. Then for each framing map $e=e_{\mathrm{in}}\oplus e_{\mathrm{out}}\oplus e_{\mathrm{m}}$, the element $\mathfrak{e}_{\mathrm{m}}\in\mathcal{A}_{\mathrm{m}}$ is realized as the map $e_{\mathrm{m}}:F_{\mathrm{m}}\to V$, and $\mathfrak{e}_{\mathrm{m}}^*$ is realized as the metric adjoint $(h(e_{\mathrm{m},l},\cdot))_{l=1}^{n_{\mathrm{m}}}:V\to F_{\mathrm{m}}=\mathbb{C}^{n_{\mathrm{m}}}$.

To consider linear maps that have the state space V as both domain and target, we can form the subalgebra

$$\mathcal{A}_{\mathrm{m},0}:=\mathcal{A}\cdot\mathcal{A}_{\mathrm{m}}\cdot\mathcal{A}$$

which is simply the algebra generated by $\mathcal{A}$ and $\mathfrak{e}_{\mathrm{m}}\mathfrak{e}_{\mathrm{m}}^*$. An element $a\in\mathcal{A}_{\mathrm{m},0}$ is understood as a linear algorithm. Fixing $(w,e)\in R$, each linear algorithm $a\in\mathcal{A}_{\mathrm{m},0}$ is associated with a linear function $f^a\colon F_{\mathrm{in}}\to F_{\mathrm{out}}$,

$$f^a(v):=e_{\mathrm{out}}^*(a\cdot e_{\mathrm{in}}(v))$$

which is called a machine function. ($e_{\mathrm{out}}^*:V\to F_{\mathrm{out}}$ is the metric adjoint $(h(e_{\mathrm{out},l},\cdot))_{l=1}^{n_{\mathrm{out}}}$.) In other words, we have the map

$$R\times\mathcal{A}_{\mathrm{m},0}\to\mathrm{Hom}(F_{\mathrm{in}},F_{\mathrm{out}})$$

which is linear in the second component.

So far, this is just a linear model. In order to capture non-linearity, we also need to incorporate *non-linear operations* $\sigma_1,\ldots,\sigma_N$. Let's define these as functions $V\to V$ for the moment. (In the next subsection, we shall see that defining in this way is not good from the moduli point of view and will thus need to modify this definition.)

Consider the $\mathbb{C}$-near-ring $\tilde{\mathcal{A}}=\mathcal{A}\{\varsigma_1,\ldots,\varsigma_N\}$. The elements ς_j are algebraic symbols for recording the non-linear operations σ_j. See Definition 3.13 for the notion of a near-ring. Essentially, it is recording the compositions of module maps and the non-linear operations. Similar to above, we take the augmented near-ring

$$\tilde{\mathcal{A}}_{\mathrm{m}}=\tilde{\mathcal{A}}\langle 1_{\mathrm{m}},\mathfrak{e}_{\mathrm{m}},\mathfrak{e}_{\mathrm{m}}^*\rangle/\tilde{I}\tag{2}$$

where $\tilde{I}$ is generated by the elements in I as in (1), together with the elements

$$\varsigma_l\cdot 1_{\mathrm{m}},\quad 1_{\mathrm{m}}\cdot\varsigma_l.$$

(Including these two extra elements means we require ς_l and 1_{m} to compose to be zero. We need these relations since σ_l is acting on V and 1_{m} is acting on F_{m}.) An element $\tilde{\gamma} \in \tilde{\mathcal{A}}_{\mathrm{m},0} := \mathcal{A} \cdot \tilde{\mathcal{A}}_{\mathrm{m}} \cdot \mathcal{A}$ is understood as a non-linear algorithm.

Fixing $(w, e) \in R$, each algorithm $\tilde{\gamma} \in \tilde{\mathcal{A}}_{\mathrm{m},0}$ is associated with a non-linear machine function $f^{\tilde{\gamma}}_{(w,e)} \colon F_{\mathrm{in}} \to F_{\mathrm{out}}$,

$$f^{\tilde{\gamma}}_{(w,e)}(v) = e^*_{\mathrm{out}}\left(\tilde{\gamma} \circ_{(w,e)} e_{\mathrm{in}}(v)\right). \tag{3}$$

That is, we have the map

$$R \times \tilde{\mathcal{A}}_{\mathrm{m},0} \to \mathrm{Map}(F_{\mathrm{in}}, F_{\mathrm{out}}).$$

Example 3.1. *We will use the example of Section 2 to illustrate. In the example, the input framing space is* $F_{\mathrm{in}} = \mathbb{C}^{784}$ *and the output framing space is* $F_{\mathrm{out}} = \mathbb{C}^{10}$. $F_m = \mathbb{C}^{2d_1}$ *where* d_1 *is the number of neurons in the hidden layer of the network. The state space* V *decomposes as* $V = V_{\mathrm{in}} \oplus V_{\mathrm{out}} \oplus V_{\mathrm{m}}$ *where* $V_{\mathrm{in}} = \bigoplus_{i=1}^{784} \mathbb{C}$, $V_{\mathrm{out}} = \bigoplus_{i=1}^{10} \mathbb{C}$, *and* $V_{\mathrm{m}} = \bigoplus_{i=1}^{d_1} \mathbb{C}$, *with each factor of* $\mathbb{C}$ *corresponding to a neuron.* V *is equipped with the standard metric.*

The input framing maps e_{in} *and* e_{out} *consist of linear functions* $(e^0_i) : \mathbb{C} \to \mathbb{C}$ *for* $i = 1, \ldots, d_0 = 784$ *and* $(e^2_k) : \mathbb{C} \to \mathbb{C}$ *for* $k = 1, \ldots, d_2 = 10$ *respectively. The map* e_{m} *consists of* $e^1_j = (e_j, b_j) : \mathbb{C}^2 \to \mathbb{C}$ *for* $j = 1, \ldots, d_1$, *where* b_j *is called the bias of the* j*-th neuron in the hidden layer.*

For the moment, we take non-linear functions $\sigma^1_j : V^1_j \to V^1_j$, *where* $V^1_j = \mathbb{C}$ *for* $j = 1, \ldots, d_1$. *In the next subsection, we will modify this point in order to make the construction well-defined over the moduli space.*

The algebra $\mathcal{A}$ *is the path algebra of the underlying directed graph of the network with concatenation as multiplication. For example, the path* a^1_{11} *from the* $(0, 1)$ *neuron in the input layer to the* $(1, 1)$ *neuron in the hidden layer can be multiplied with the path* a^2_{21} *from the* $(1, 1)$ *neuron to the* $(2, 2)$ *neuron to get* $a^2_{21}a^1_{11}$, *which is a path from the* $(0, 1)$ *neuron to the* $(2, 2)$ *neuron. On the other hand,* $a^1_{11}a^2_{21}$ *is defined as zero since the concatenation is not valid.*

In this example, the typical algorithm is

$$\tilde{\gamma} = \sum_{k=1}^{10} ((\mathfrak{e}_{out}))^* \sum_{j=1}^{n} a^1_{kj}\varsigma^1_j \circ \left(\sum_{i=1}^{784} a^0_{ji}e^0_i + b_j\right)$$

resulting in the non-linear machine function

$$(f^{\tilde{\gamma}}_{(w^1,w^2,b)}(x))_k = \sum_{j=1}^{n} w^2_{kj}\sigma^1_j\left(\sum_{i=1}^{784} w^1_{ji} \cdot x_i + b_j\right).$$

3.2. Construction over moduli spaces

A guiding principle in mathematics and physics is that isomorphic objects should produce the same result. In other words, we want to have $f^{\tilde{\gamma}}_{(w,e)}$ well-defined over the moduli stack of framed $\mathcal{A}$-modules $\mathcal{M} = [R/G]$ for $G = \mathrm{GL}(V)$. Let's recall the following definition.

Definition 3.2. *For two framed $\mathcal{A}$-modules (V, w, e) and (V', w', e'), where both e and e' have the same domain F, a morphism (or an isomorphism) from (V, w, e) to (V', w', e') is a linear map (or a linear isomorphism) $g : V \to V'$ such that $w'(a) \circ g = g \circ w(a)$ for all $a \in \mathcal{A}$ and $e' = g \circ e$.*

Ideally, we would want to extend this definition of morphism to the $\mathbb{C}$-near-rings $\widetilde{\mathcal{A}}$ we constructed in Section 3.1. Unfortunately, we cannot directly do this due to the additional data of the non-linear functions $\sigma : V \to V$. Any useful non-linear function $\sigma : V \to V$ cannot satisfy $\mathrm{GL}(V)$-equivariance:

$$g \cdot (\sigma(v)) = \sigma(g \cdot v) \text{ for all } g \in \mathrm{GL}(V). \tag{4}$$

It produces a crucial gap between the subject of machine learning and representation theory.

Here is a simple solution to this problem. Let $\mathcal{V}$ be the universal bundle over the moduli stack $\mathcal{M}$, which is descended from the trivial bundle $V \times R$, where $G = \mathrm{GL}(V)$ acts diagonally.

Rather than defining σ as a single linear map $V \to V$, let's take σ to be a fiber-bundle map $V \times R \to V \times R$ over R. Then σ descends as a fiber-bundle map $\mathcal{V} \to \mathcal{V}$ over $\mathcal{M}$ if it satisfies the equivariance equation

$$g \cdot \left(\sigma_{(w,e)}(v)\right) = \sigma_{(g\cdot w, g\cdot e)}(g \cdot v) \text{ for all } g \in \mathrm{GL}(V). \tag{5}$$

The difference between Equation (5) and (4) is that σ is now allowed to also depend on $(w, e) \in R$.

Now suppose we have $\mathrm{GL}(V)$-equivariant fiber-bundle maps $\sigma_1, \ldots, \sigma_N : V \times R \to V \times R$. As in the last subsection, we have the map $R \times \tilde{\mathcal{A}}_{\mathrm{m},0} \to \mathrm{Map}(F_{\mathrm{in}}, F_{\mathrm{out}})$ by realizing $\varsigma_i \in \tilde{\mathcal{A}}$ as $(\sigma_i)_{(w,e)} : \mathcal{V} \to \mathcal{V}$.

Recall that we have used a Hermitian metric on V for taking the adjoint of framing e^*. To make sure e^* is also equivariant, we need to equip V with a family of Hermitian metrics $h_{(w,e)}$ for $(w, e) \in R$, in a $\mathrm{GL}(V)$-equivariant way:

$$h_{(g\cdot w, g\cdot e)}(g \cdot u, g \cdot v) = h_{(w,e)}(u, v) \text{ for all } g \in \mathrm{GL}(V). \tag{6}$$

That is, h descends to be a Hermitian metric on the universal bundle $\mathcal{V}$ over $\mathcal{M}$.

Note that we are NOT asking for $\mathrm{GL}(V)$-invariance $h(g \cdot u, g \cdot v) = h(u, v)$ for a single metric h, which is impossible except for in the trivial case.

Later in Section 4 we will discuss these families of $\mathrm{GL}(V)$-equivariant Hermititian metrics in more detail. In particular, we will construct them for the case that $[R/\mathrm{GL}(V)]$ is a framed quiver moduli space.

Proposition 3.3. *In the above setting, the non-linear machine function defined by Equation (3) satisfies the equivariance* $f^{\tilde{\gamma}}_{(w,e)} = f^{\tilde{\gamma}}_{g\cdot(w,e)}$ *for all* $g \in \mathrm{GL}(V)$.

Proof. The fiber-bundle map $f^{\tilde{\gamma}}_{(w,e)} : V \times R \to V \times R$ defined by (3) is a composition of $e^*_{\mathrm{out}} = (h_{(w,e)}(e_{\mathrm{out},l}, \cdot))_{l=1}^{n_{\mathrm{out}}}$, w_a for $a \in \mathcal{A}$, the fiber-bundle maps $(\sigma_i)_{(w,e)} : V \times R \to V \times R$, and e_{in}. Under the action of $g \in \mathrm{GL}(V)$, They change to

$$e^*_{\mathrm{out}} = (h_{g\cdot(w,e)}(g \cdot e_{\mathrm{out},l}, \cdot))_{l=1}^{n_{\mathrm{out}}} = (h_{(w,e)}(e_{\mathrm{out},l}, g^{-1}(\cdot)))_{l=1}^{n_{\mathrm{out}}} = e^*_{\mathrm{out}} \cdot g^{-1},$$

$g \cdot w_a \cdot g^{-1}$,

$$\sigma_{(g\cdot w, g\cdot e)} = g \cdot \sigma_{(w,e)}(g^{-1}(\cdot))$$

and $g \cdot e_{\mathrm{in}}$ respectively, using Equation (5) and (6). The composition remains the same. □

In this way, we obtain the map $\mathcal{M} \times \tilde{\mathcal{A}}_{\mathrm{m},0} \to \mathrm{Map}(F_{\mathrm{in}}, F_{\mathrm{out}})$.

In applications, we need concrete fiber bundle maps $\sigma : \mathcal{V} \to \mathcal{V}$. They can be constructed using the Hermitian metric h on $\mathcal{V}$ as follows. Given any function $\sigma^F \colon F_{\mathrm{m}} \to F_{\mathrm{m}}$, define $\sigma_{(w,e)}$ as

$$\sigma_{(w,e)}(v) \coloneqq e^{(\mathrm{m})} \cdot \sigma^F \left(h_{(w,e)} \left(e_1^{(\mathrm{m})}, v \right), \ldots, h_{(w,e)} \left(e_{n_{\mathrm{m}}}^{(\mathrm{m})}, v \right) \right).$$

In other words, we observe and record the state v to memory using $e^{(\mathrm{m})}$ and h; then we perform the non-linear operation σ^F on the memory F_{m}; finally we send it back as a state in V. Unlike the setting in the last subsection, the non-linear operation σ^F is now defined on the framing space F_{m} instead of on the basis-free state space V.

Proposition 3.4. *The above* $\sigma_{(w,e)} \colon V \times R \to V \times R$ *is* $\mathrm{GL}(V)$*-equivariant.*

Proof.

$\sigma_{(g\cdot w, g\cdot e)}(g \cdot v)$

$$= g \cdot e^{(\mathrm{m})} \cdot \sigma^F \left(h_{(g\cdot w, g\cdot e)} \left(g \cdot e_1^{(\mathrm{m})}, g \cdot v \right), \ldots, h_{(g\cdot w, g\cdot e)} \left(g \cdot e_{n_\mathrm{m}}^{(\mathrm{m})}, g \cdot v \right) \right)$$
$$= g \cdot e^{(\mathrm{m})} \cdot \sigma^F \left(h_{(w,e)} \left(e_1^{(\mathrm{m})}, v \right), \ldots, h_{(w,e)} \left(e_{n_\mathrm{m}}^{(\mathrm{m})}, v \right) \right) = g \cdot \sigma_{(w,e)}(v)$$

using Equation (6). □

The non-linear operations are called activation functions in machine learning. We conclude the current setting by the following definition.

Definition 3.5. *An* activation module *consists of:*

1. *a (noncommutative) algebra* $\mathcal{A}$ *and vector spaces* V, $F = F_\mathrm{in} \oplus F_\mathrm{out} \oplus F_\mathrm{m}$*;*
2. *A family of metrics* $h_{(w,e)}$ *on* V *over the space of framed* $\mathcal{A}$*-modules*

$$R = \mathrm{Hom}_\mathrm{alg}(\mathcal{A}, \mathrm{End}(V)) \times \mathrm{Hom}(F, V)$$

which is $\mathrm{GL}(V)$*-equivariant;*
3. *a collection of possibly non-linear functions*

$$\sigma_j^F : F_\mathrm{m} \to F_\mathrm{m}.$$

The data of (1) and (2) (without (3)) is called a Hermitian family of framed modules.

Figure 3 shows a schematic picture of an activation module.

In this setting, σ_j^F is a function on F_m. We take the subalgebra

$$\mathcal{L}(\mathcal{A}_\mathrm{m}) := \mathfrak{e}_\mathrm{m}^* \cdot \mathcal{A}_\mathrm{m} \cdot \mathfrak{e}_\mathrm{m} \tag{7}$$

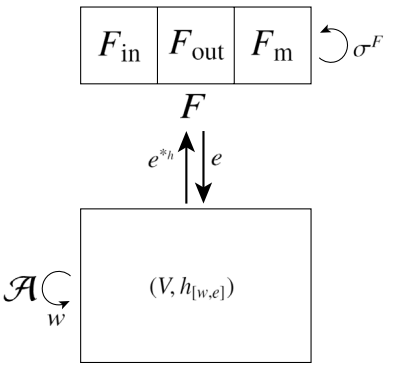

Figure 3

consisting of loops at F_{m}, the near-ring

$$\widetilde{\mathcal{A}_{\mathrm{m}}} := (\mathcal{L}(\mathcal{A}_{\mathrm{m}}))\{\varsigma_1, \ldots, \varsigma_N\}, \tag{8}$$

and

$$(\widetilde{\mathcal{A}_{\mathrm{m}}})_0 := \mathcal{A} \cdot \mathfrak{e}_{\mathrm{m}} \cdot \widetilde{\mathcal{A}_{\mathrm{m}}} \cdot \mathfrak{e}_{\mathrm{m}}^* \cdot \mathcal{A}.$$

Note that $\widetilde{\mathcal{A}_{\mathrm{m}}}$ is different from $\tilde{\mathcal{A}}_{\mathrm{m}}$ in Equation (2), since we now have non-linear functions defined on F instead of V.

Using Proposition 3.3 and 3.4, each algorithm $\tilde{\gamma} \in (\widetilde{\mathcal{A}_{\mathrm{m}}})_0$ and $[w, e] \in \mathcal{M}$ gives a machine function $f^{\tilde{\gamma}}_{[w,e]}$. This gives a map

$$(\widetilde{\mathcal{A}_{\mathrm{m}}})_0 \to \Gamma\left(\mathcal{M}, \mathbf{Map}\left(F_{\mathrm{in}}, F_{\mathrm{out}}\right)\right).$$

In applications, an activation module may consist of several linear submodules, which are connected by possibly non-linear transitions σ_i^F. This means the algebra $\mathcal{A}$ is a direct sum $\bigoplus_{k\in K} \mathcal{A}^{(k)}$ where each $\mathcal{A}^{(k)}$ is understood as a linear component of the activation module (and K is an index set). Similarly, we have $V = \bigoplus_{k\in K} V^{(k)}$ and $F = \bigoplus_{k\in K} F^{(k)}$. We take the moduli stack $\prod_{k\in K}[R^{(k)}/\mathrm{GL}(V^{(k)})]$ (where $R^{(k)} = \mathrm{Hom}_{\mathrm{alg}}\left(A^{(k)}, \mathrm{End}\left(V^{(k)}\right)\right) \times \mathrm{Hom}\left(F^{(k)}, V^{(k)}\right)$) instead of $[R/\mathrm{GL}(V)]$. Each $F^{(k)}$ has three components $F^{(k)} = F^{(k)}_{\mathrm{in}} \oplus F^{(k)}_{\mathrm{out}} \oplus F^{(k)}_{\mathrm{m}}$ (where some of the components can simply be $\{0\}$). Furthermore, the non-linear functions $\sigma_j^F : F_{\mathrm{m}} \to F_{\mathrm{m}}$ is a composition $\iota \circ s_j^F \circ \pi$, where $s_j^F : F_{\mathrm{m}}^{(p_{j,1})} \times \ldots \times F_{\mathrm{m}}^{\left(p_{j,m_j}\right)} \to F_{\mathrm{m}}^{(q_{j,1})} \times \ldots \times F_{\mathrm{m}}^{\left(q_{j,n_j}\right)}$ for some fixed $\{p_{j,1}, \ldots, p_{j,m_j}\}$ and $\{q_{j,1}, \ldots, q_{j,n_j}\}$; π is the projection $F_{\mathrm{m}} \to F_{\mathrm{m}}^{(p_{j,1})} \oplus \ldots \oplus F_{\mathrm{m}}^{\left(p_{j,m_j}\right)}$ and ι is the inclusion (or extension by zero) $F_{\mathrm{m}}^{(q_{j,1})} \oplus \ldots \oplus F_{\mathrm{m}}^{\left(q_{j,n_j}\right)} \to F_{\mathrm{m}}$. Finally, h is a direct sum $h_{(w,e)} = \bigoplus_{k\in K} h_{\left(w^{(k)}, e^{(k)}\right)}$ where each $h_{\left(w^{(k)}, e^{(k)}\right)}$ is a family of metrics $h_{\left(w^{(k)}, e^{(k)}\right)}$ on $V^{(k)}$ over the space of framed $A^{(k)}$-modules $R^{(k)}$ which is $\mathrm{GL}\left(V^{(k)}\right)$-equivariant.

We can also define a closely related setting that uses unitary framed modules, which takes the unitary group $U(V, h)$ in place of $\mathrm{GL}(V)$, and takes a single Hermitian metric h in place of a family of Hermitian metrics.

Definition 3.6. *A* unitary activation module *consists of:*

1. *A Hermitian vector space (V, h), a framing vector space $F = F_{\mathrm{in}} \oplus F_{\mathrm{out}} \oplus F_{\mathrm{m}} = \mathbb{C}^n$ (equipped with the standard metric), and unitary framing maps $e_\bullet : F_\bullet \to V$, where $\bullet = \mathrm{in}, \mathrm{out}, \mathrm{m}$.*

2. *A group ring* $\mathcal{A} = \mathbb{C}[G]$ *where* G *is a subgroup of the unitary group* $U(V,h)$. $\mathbb{C}[G]$ *consists of linear combinations* $\sum_{g\in G} c_g g$ *for* $c_g \in \mathbb{C}$.
3. *a collection of possibly non-linear functions*

$$\sigma_j^F : F_{\mathrm{m}} \to F_{\mathrm{m}}.$$

Such a setting is well suited for quantum computing. Namely, (V,h) can be taken to be the state space of a quantum system of particles. G is a subgroup of unitary operators on (V,h). F_{m} can be taken to have the same dimension as V, and $e_{\mathrm{m}} : F_{\mathrm{m}} \to V$ maps the standard basis of F_{m} to an assigned unitary basis of V. (For instance, the assigned basis can be $\{|00\rangle, |01\rangle, |10\rangle, |11\rangle\}$ for a 2-qubit system). There is a probabilistic projection $\sigma_0 : F_{\mathrm{m}} \to F_{\mathrm{m}}$ that corresponds to wave-function collapse following each observation. We also have other non-linear classical operations σ_j^F on F_{m}.

In application, we are given input data $v \in F_{\mathrm{in}}$. This v (normalized to have length 1) is sent to the Hermitian state space V by e_{in}, and operated under a prescribed linear algorithm $a \in \mathbb{C}[G]$. Then the system is observed and recorded using the basis e_{m}. This gives $\sigma_0 \cdot \sum_l h(e_{\mathrm{m},l}, a \cdot e_{\mathrm{in}} \cdot v) e_{\mathrm{m},l}$. The recorded memory can be operated by a non-linear algorithm consisting of σ_j^F. The process can be iterated and give a function $F_{\mathrm{in}} \to F_{\mathrm{out}}$.

In this paper, we focus on Definition 3.5, for the purpose of neural networks and deep learning which works with $\mathrm{GL}(V)$ rather than $U(V)$.

3.3. Noncommutative geometry and machine learning

We have formulated a computing machine by a Hermitian family of framed $\mathcal{A}$-modules and a collection of non-linear functions. If we ignore the non-linear functions for the moment, and merely consider the augmented algebra $\mathcal{A}_{\mathrm{m}}$, it fits well to the framework of noncommutative geometry developed by Connes [11], Cuntz-Quillen [12], Ginzburg [18]. Below we give a quick review and apply to our situation. [32] gives a beautiful survey on this theory. We will extend it to near-ring in the next subsection.

3.3.1. A quick review The theory develops an analog of the de Rham complex of differential forms for an associative algebra A over a field $\mathbb{K}$ (that we take to be $\mathbb{C}$ in this paper). This is a crucial step to develop the notions of cohomology, connection and curvature for the noncommutative space associated to A and its associated vector bundles.

The noncommutative differential forms can be described as follows. Consider the quotient vector space $\overline{A} = A/\mathbb{K}$ (which is no longer an algebra). We

think of elements in $\overline{A}$ as differentials. Define

$$D(A) := \bigoplus_{n\in\mathbb{Z}_{\geq 0}} D(A)_n,\ D(A)_n := A\otimes\overline{A}\otimes\ldots\otimes\overline{A}$$

where n copies of $\overline{A}$ appear in $D(A)_n$, and the tensor product is over the ground field $\mathbb{K}$. We should think of elements in $\overline{A}$ as *matrix-valued* differential one-forms. Note that $X\wedge X$ may not be zero, and $X\wedge Y\neq -Y\wedge X$ in general for matrix-valued differential forms X, Y.

The differential $d_n : D(A)_n \to D(A)_{n+1}$ is defined as

$$d_n(a_0\otimes\overline{a_1}\otimes\ldots\otimes\overline{a_n}) := 1\otimes\overline{a_0}\otimes\ldots\otimes\overline{a_n}.$$

The product $D(A)_n\otimes D(A)_{m-1-n}\to D(A)_{m-1}$ is more tricky:

$$\begin{aligned}&(a_0\otimes\overline{a_1}\otimes\ldots\otimes\overline{a_n})\cdot(a_{n+1}\otimes\overline{a_{n+2}}\otimes\ldots\otimes\overline{a_m})\\ &:=(-1)^n a_0a_1\otimes\overline{a_2}\otimes\ldots\otimes\overline{a_m}+\sum_{i=1}^{n}(-1)^{n-i}a_0\otimes\overline{a_1}\otimes\ldots\otimes\overline{a_ia_{i+1}}\otimes\ldots\otimes\overline{a_m}\end{aligned}\tag{9}$$

which can be understood by applying the Leibniz rule on the terms $\overline{a_ia_{i+1}}$. Note that we have chosen representatives $a_i\in A$ for $i=1,\ldots,n+1$ on the RHS, but the sum is independent of choice of representatives (while the product $\overline{a_ia_{i+1}}$ itself depends on representatives).

The above product in particular gives a bimodule structure on $D(A)$ over $A=D(A)_0$. For instance, $D(A)_1$ has the bimodule structure

$$a\cdot(a_0\otimes\overline{a_1})=aa_0\otimes\overline{a_1},\ (a_0\otimes\overline{a_1})\cdot a=-a_0a_1\otimes\overline{a}+a_0\otimes\overline{a_1a}.$$

(If a_1 is replaced by a_1+k for $k\in\mathbb{K}$, then RHS $=-a_0a_1\otimes\overline{a}-ka_0\otimes\overline{a}+a_0\otimes\overline{a_1a}+ka_0\otimes\overline{a}=-a_0a_1\otimes\overline{a}+a_0\otimes\overline{a_1a}$ remains unchanged.)

By [12],

$$d^2=0.$$

The above differential d and product defines a dg-algebra structure on $D(A)$; indeed this is the unique one that satisfies $a_0\cdot da_1\cdot\ldots\cdot da_n=a_0\otimes\overline{a_1}\otimes\ldots\otimes\overline{a_n}$. Moreover, $(D(A),i)$, where $i: A\to D(A)_0=A$ is the identity map, has the following universal property: for every (Γ,ψ) where Γ is a dg algebra and $\psi: A\to\Gamma_0$ is an algebra homomorphism, there exists an extension as a dg-algebra map $u_\psi: D(A)\to\Gamma$ such that the degree-zero part satisfies $(u_\psi)_0\circ i=\psi$.

Here is another realization of differential forms for A. First, define the A-bimodule $\Omega^1(A) := \mathrm{Ker}\,(\mu)$ where $\mu : A \otimes A \to A$ is the multiplication map for A. Moreover, define $d : A \to \Omega^1(A)$ by $da := 1 \otimes a - a \otimes 1$. Thus $\sum_i a_i da_i'$ for $a_i, a_i' \in A$ is an element in $\Omega^1(A)$. Conversely, any element in $\Omega^1(A)$ is of the form $\sum_i a_i \otimes a_i'$ with $\sum_i a_i \cdot a_i' = 0$, and this is equal to

$$\sum_i a_i da_i' = -\sum_i (da_i) a_i'.$$

Then we take the tensor algebra

$$\Omega^\bullet(A) := T_A(\Omega^1(A)) = \bigoplus_{i \in \mathbb{Z}_{\geq 0}} \Omega^1(A) \otimes_A \ldots \otimes_A \Omega^1(A)$$

where there are i copies of $\Omega^1(A)$ for the summands on the right. An element in $\Omega^\bullet(A)$ takes the form $a_1 db_1 \otimes_A a_2 db_2 \otimes_A \ldots \otimes_A a_k db_k \cdot a_{k+1}$. Recall that tensoring over A means the identification $db_1 \cdot a \otimes_A db_2 = db_1 \otimes_A a db_2$.

The two defined graded algebras $\Omega^\bullet(A)$ and $D(A)$ are isomorphic. For one forms, we have the A-bimodule map $\psi : \Omega^1(A) \to D(A)_1$ defined by $da \mapsto 1 \otimes \overline{a}$. It has the inverse $a_0 \otimes \overline{a_1} \mapsto a_0 \otimes a_1 - a_0 a_1 \otimes 1$ (which is again independent of choice of representative a_1). For higher forms, $\Omega^n \to D(A)_n$ is given by $\alpha_1 \otimes_A \ldots \otimes_A \alpha_n \mapsto \psi(\alpha_1) \cdot \ldots \cdot \psi(\alpha_n)$ (where the non-trivial product on $D(A)$ is given in Equation (9)), whose inverse is $a_0 \otimes \overline{a_1} \otimes \ldots \otimes \overline{a_n} = (a_0 \otimes \overline{a_1}) \cdot (1 \otimes \overline{a_2}) \ldots (1 \otimes \overline{a_n}) \mapsto \psi^{-1}(a_0 \otimes \overline{a_1}) \otimes_A \psi^{-1}(1 \otimes \overline{a_2}) \otimes_A \ldots \otimes_A \psi^{-1}(1 \otimes \overline{a_n})$.

The Karoubi-de Rham complex is defined as

$$DR^\bullet(A) := \Omega^\bullet(A)/[\Omega^\bullet(A), \Omega^\bullet(A)] \tag{10}$$

where $[a, b] := ab - (-1)^{ij} ba$ is the graded commutator for a graded algebra. d descends to be a well-defined differential on $DR^\bullet(A)$. Note that $DR^\bullet(A)$ is not an algebra since $[\Omega^\bullet(A), \Omega^\bullet(A)]$ is not an ideal. $DR^\bullet(A)$ is the noncommutative analog for the space of de Rham forms. Moreover, there is a natural map by taking trace to the space of G-invariant differential forms on the space of representations $R(A)$:

$$DR^\bullet(A) \to \Omega^\bullet(R(A))^G. \tag{11}$$

The subspaces $DR^0(A)$ and $DR^1(A)$ will be the most relevant to us. We have $DR^0(A) = A/[A, A]$ and $DR^1(A) = \Omega^1(A)/[A, \Omega^1(A)]$.

Dually, derivations $\theta \in \mathrm{Der}(A)$ play the role of vector fields. A derivation $\delta : A \to A$ is a linear map satisfying $\delta(ab) = \delta(a) \cdot b + a \cdot \delta(b)$. $\mathrm{Der}(A)$ is the

vector space of all derivations. We have the A-bimodule map $\iota_\theta : \Omega^1(A) \to A$, $\iota_\theta(da) := \theta(a)$ called contraction. ι_θ extends to $\Omega^\bullet(A) \to \Omega^{\bullet-1}(A)$ by using graded Leibniz rule, and descends to $DR^\bullet(A) \to DR^{\bullet-1}(A)$.

The following version of differential forms relative to a subalgebra [12] will be useful for framings and quivers. Let $B \subset A$ be a commutative subalgebra. We take

$$D(A/B)_n := A \otimes_B \bar{A} \otimes_B \ldots \otimes_B \bar{A}$$

where $\bar{A}$ is the vector space

$$\bar{A} := A/B.$$

Then we repeat the same definitions as above for $DR^\bullet(A/B)$. Note that zero-th forms are the same as before: $DR^0\bullet(A/B) = DR^0\bullet(A)$. There is a natural map [18, 3]

$$DR^\bullet(A/B) \to \Omega^\bullet\left(R_B\left(A\right)\right)^{G_B}$$

where $R_B(A)$ is the set of A-modules whose restriction to B is equal to a prescribed B-module, and G_B is the subgroup in $\mathrm{GL}(V)$ that preserves the prescribed B-bimodule structure.

In the context of A being the path algebra of a quiver, we shall take B to be the subalgebra generated by the trivial paths 1_i at all vertices $i \in Q_0$. Then a differential form

$$a_0(da_1)(da_2)\ldots(da_k) \in DR^\bullet(A/B)$$

is non-zero only if the paths a_i can be concatenated: $t(a_j) = h(a_{j+1})$ for all $j \in \mathbb{Z}/(k+1)$. In this case, a prescribed B-module structure on V is given by a decomposition $V = \bigoplus_{i\in Q_0} V_i$ and 1_i acts as the projection $V \to V_i$. Then $G_B = \prod_{i\in Q_0} \mathrm{GL}(V_i)$.

3.3.2. Connection to linear machine learning Now we come back to the context of the last subsection. The additional ingredient we need to take care of is the equivariant family of Hermitian metrics h on the A-modules.

To precisely match the language, first let's modify the definition for $\mathcal{A}_\mathrm{m}$ (Equation (1)) as follows. Recall that the framing vector space $F = F_\mathrm{in} \oplus F_\mathrm{out} \oplus F_\mathrm{m} = \mathbb{C}^{n_\mathrm{in}} \oplus \mathbb{C}^{n_\mathrm{out}} \oplus \mathbb{C}^{n_\mathrm{m}}$, where $\dim F = n$. Then a framing e can be written as $(e_1 \ldots e_n)$ where $e_j \in V$, and e^* is the column vector $(e_1^*, \ldots, e_n^*)$ where $e_j^* \in V^*$.

First, we take the augmentation

$$\mathcal{A}^\mathfrak{e} := \mathcal{A}\langle 1_F, \mathfrak{e}_j : j = 1, \ldots, n\rangle / I$$

where I is the two-sided ideal generated by $1_F \cdot \mathfrak{e}_j$, $\mathfrak{e}_j \cdot 1_F - \mathfrak{e}_j$, $1_{\mathcal{A}} \cdot \mathfrak{e}_j - \mathfrak{e}_j$, $\mathfrak{e}_j \mathfrak{e}_k$, $\mathfrak{e}_j \cdot a, a \cdot 1_F$, $1_F \cdot a$ for all $a \in \mathcal{A}, j, k = 1, \ldots, n$.

Then we take its *doubling* $\hat{\mathcal{A}}$, which is generated by two copies of $\mathcal{A}^{\mathfrak{e}}$ (whose generators are denoted by $a, 1_F, \mathfrak{e}_j$ and $a^*, 1_F^*, \mathfrak{e}_j^*$ respectively), quotient out the ideal of relations $1_{\mathcal{A}} - 1_{\mathcal{A}}^*, 1_F - 1_F^*$. The unit of $\hat{\mathcal{A}}$ is

$$1_{\hat{\mathcal{A}}} = 1_F + 1_{\mathcal{A}}.$$

We also use the rule $(ab)^* := b^* a^*$ to define the formal adjoint of a general element in $\hat{\mathcal{A}}$.

Remark 3.7. *This doubling procedure is standard in the construction of Nakajima quiver varieties, which is an algebraic analog of taking the cotangent bundle (or complexification) of a variety. We will restrict to a section to go back to $[R/G]$.*

In the notation of the last subsection, we take $A = \hat{\mathcal{A}}$ and the commutative subalgebra

$$B = \mathrm{Span}_{\mathbb{C}}\{1_F, 1_{\mathcal{A}}\} \subset \hat{\mathcal{A}}.$$

Consider $V \oplus \mathbb{C}$. We fix its B-module structure in the way that $1_{\mathcal{A}}$ and 1_F act as $(\mathrm{Id}_V, 0)$ and $(0, \mathrm{Id}_{\mathbb{C}})$ respectively. $V \oplus \mathbb{C}$ can be equipped with an $\hat{\mathcal{A}}$-module structure that restricts to be this fixed B-module structure.

Lemma 3.8. *Given a Hermitian family of framed modules $(\mathcal{A}, V, F, h)$, there is a one-to-one correspondence between elements in*

$$R = \mathrm{Hom}_{\mathrm{alg}}(\mathcal{A}, \mathrm{End}(V)) \times \mathrm{Hom}(F, V)$$

and $\hat{\mathcal{A}}$-modules of the form $V \oplus \mathbb{C}$ that respect the B-module structure and have $\mathfrak{e}_j^*, a^*$ acting as the adjoints of e_j and $w(a)$ respectively with respect to h.

Proof. Given $(w, e) \in R$, the $\hat{\mathcal{A}}$-module structure on $V \oplus \mathbb{C}$ is defined as follows. The action of $\mathcal{A}$ on V is given by w, and $\mathcal{A}$ acts on the component $\mathbb{C}$ by zero. The element $\mathfrak{e}_j$ acts as the linear map $e_j : \mathbb{C} \to V$ where e_j is the j-th column of e, and acts on V trivially. $\mathfrak{e}_j^*$ and a^* act on the component $\mathbb{C}$ by zero, and act as the adjoint maps of $\mathfrak{e}_j$ and $w(a)$ with respect to h. The adjoint maps are

$$e_j^{*_h} : V \to \mathbb{C},\ e_j^{*_h}(v) = h(e_j, v)$$

and

$$w(a)^{*_h} = h_{(w,e)}^{-1} w(a)^* h_{(w,e)}$$

in matrix form.

Conversely, since the $\hat{A}$-module is required to restrict as the given B-module structure, we must have $\mathcal{A}$ acting trivially on the component $\mathbb{C}$, $\mathfrak{e}_j$ acting trivially on V, and $\mathfrak{e}_j^*$ acting trivially on $\mathbb{C}$. (For instance, $a = a \cdot 1_{\mathcal{A}}$ acts as $(a, 0)$ on $V \oplus \mathbb{C}$.) Then the action of $\mathcal{A}$ and $(\mathfrak{e}_j : j = 1, \ldots, n)$ gives an element in R. □

Similar to (11), we have the following map for $\hat{\mathcal{A}}$. The only difference is that for the forms $d\mathfrak{e}_j^*$ and da^*, the corresponding forms on

$$R = \mathrm{Hom}_{\mathrm{alg}}(\mathcal{A}, \mathrm{End}(V)) \times \mathrm{Hom}(F, V)$$

are defined using the metrics h.

Proposition 3.9. *Given a Hermitian family of framed modules $(\mathcal{A}, V, F, h)$, there is a (degree-preserving) map*

$$DR^\bullet(\hat{\mathcal{A}}/B) \to \Omega^\bullet(R)^{G_B}$$

that commutes with the differential, and is equal to the trace of the corresponding representations given in Lemma 3.8 when restricted to $DR^0(\hat{\mathcal{A}}/B) \to \Omega^0(R)^{G_B}$.

Proof. $DR^\bullet(\hat{\mathcal{A}}/B)$ is generated by the one forms da, da^*, $d\mathfrak{e}_j$ and $d\mathfrak{e}_j^*$ over $\hat{\mathcal{A}}$. For da and $d\mathfrak{e}_j$, the corresponding matrix-valued one-forms on R are obvious (by substituting a and $\mathfrak{e}_j$ by the corresponding representing matrices $w(a)$ and e_j). For $d\mathfrak{e}_j^*$ and da^*, the corresponding matrix-valued one-forms over R are

$$(\bar{\partial} e_j^*) \cdot h + e_j^* \cdot dh = (\bar{\partial} e_j^*) \cdot h + e_j^* \cdot (\bar{\partial} h + \partial h)$$

and

$$-h^{-1} \cdot dh \cdot h^{-1} w_a^* h + h^{-1}(\bar{\partial} w_a^*) h + h^{-1} w_a^* dh \tag{12}$$

respectively, where h is now represented by a square matrix in a basis of V, and e_j^* (a row vector) and w_a^* are the conjugate transpose of e_j and w_a respectively. Note that $h_{(w,e)}$ is a function on $(w, e) \in R$ and so it has a non-trivial differential dh. More intrinsically, $d\mathfrak{e}_j^*$ corresponds to $h(\nabla e_j, \cdot) + h(e_j, \nabla \cdot)$, where ∇ is the Chern connection of h on the trivial vector bundle $V \times R$ (and e_j is a section).

Note that non-zero elements in $DR^\bullet(\hat{\mathcal{A}}/B)$ are represented by loops (meaning that the source and target are the same), due to the defining equation (10).

The corresponding forms on R are obtained by composing the above matrices and taking trace. In particular, it is the trace of the corresponding representing matrix when restricted to $DR^0(\hat{\mathcal{A}}/B)$. Since trace is independent of cyclic permutations of the composition, the map $DR^\bullet(\hat{\mathcal{A}}/B) \to \Omega^\bullet(R)$ is well-defined. Moreover, it commutes with the differential by definition.

Under the action of $g \in \mathrm{GL}(V)$, $d(w(a)) \mapsto g \cdot d(w(a)) \cdot g^{-1}$, $de_j \mapsto g \cdot de_j$,

$$\begin{aligned}(\bar{\partial} e_j^*) \cdot h + e_j^* \cdot dh &\mapsto (\bar{\partial} e_j^*) g^* \cdot (g^*)^{-1} h g^{-1} + e_j^* g^* \cdot (g^*)^{-1} dh\, g^{-1} \\ &= ((\bar{\partial} e_j^*) \cdot h + e_j^* \cdot dh) \cdot g^{-1}\end{aligned}$$

and (12) transforms by $g\,(\cdot)\,g^{-1}$, using the $\mathrm{GL}(V)$-equivariance of the family of metrics h. Since trace is invariant under conjugation, the corresponding forms on R are G_B-invariant. Here $G_B = \mathbb{C}^\times \times \mathrm{GL}(V)$, where $\mathbb{C}^\times$ is Abelian and acts trivially on R. □

Remark 3.10. *Since the above uses the family of Hermitian metrics h, the resulting forms in $\Omega^\bullet(R)^{G_B}$ are no longer holomorphic. In the usual algebraic construction, we have a map ρ from $DR^p(\hat{\mathcal{A}}/B)$ to $\mathrm{GL}(V)$-invariant holomorphic $(p,0)$-forms on*

$$\left(\mathrm{Hom}_{\mathrm{alg}}\left(\mathcal{A}, \mathrm{End}\,(V)\right)\right)^2 \times \mathrm{Hom}\,(F,V) \times \mathrm{Hom}\,(V,F)\,.$$

The above can be understood as a composition of the usual map

$$\rho : DR^\bullet(\hat{\mathcal{A}}/B) \to \Omega^\bullet(R \times (\mathrm{Hom}_{\mathrm{alg}}\left(\mathcal{A}, \mathrm{End}\,(V)\right) \times \mathrm{Hom}\,(V,F)))$$

together with pulling back by the smooth section of $R \times (\mathrm{Hom}_{\mathrm{alg}}\left(\mathcal{A}, \mathrm{End}\,(V)\right) \times \mathrm{Hom}\,(V,F)) \to R$ defined by

$$e'_j = h_{(w,e)}(e_j, \cdot) = e_j^* \cdot h_{(w,e)},\ \ w'_a = h_{(w,e)}^{-1} w_a^* h_{(w,e)}.$$

On the LHS, $(e'_j : j = 1, \ldots, n) \in \mathrm{Hom}\,(V,F)$ and $w'_a \in \mathrm{End}(V)$ denotes fiber coordinates; on the RHS, e_j^* is the conjugate transpose of the column vector e_j in $(e_1 \ldots e_n) \in \mathrm{Hom}(F,V)$. Note that the action of $\mathrm{GL}(V)$ on both sides of the first and second equations are right multiplication by g^{-1} and conjugation $g\,(\cdot)\,g^{-1}$ respectively.

Now define the subalgebra

$$\mathcal{L}(\hat{\mathcal{A}}) := \bigoplus_{j,k=1}^{n} \mathfrak{e}_j^* \cdot \hat{\mathcal{A}} \cdot \mathfrak{e}_k.$$

Recall that elements in $\mathcal{L}(\hat{\mathcal{A}})$ are understood as linear algorithms.

In $DR^0(\hat{\mathcal{A}}/B) = \hat{\mathcal{A}}/(B + [\hat{\mathcal{A}}, \hat{\mathcal{A}}])$ (vector-space quotient), note that elements that do not form loops (for instance, $a \cdot \mathfrak{e}_j$ and $\mathfrak{e}_j^* \cdot a$) are in the zero class. Moreover, loops that are cyclic permutation of each other are identified as the same class.

In our context, elements in $\mathcal{L}(\hat{\mathcal{A}})$ are loops, and they descend to non-trivial elements in $DR^0(\hat{\mathcal{A}}/B)$. As a consequence:

Corollary 3.11. *An element in $\mathcal{L}(\hat{\mathcal{A}})$ induces a G-invariant function f on R where $G = \mathrm{GL}(V)$. Its differential lies in $DR^1(\hat{\mathcal{A}}/B)$ and induces the corresponding differential $df \in \Omega^1(R)^G$.*

Note that the target of $\mathfrak{e}_j^*$ and the domain of $\mathfrak{e}_j$ are the one-dimensional vector space $\mathbb{C}$. Thus the matrix corresponding to $\mathfrak{e}_j^* \cdot a \cdot \mathfrak{e}_k \in \mathcal{L}(\hat{\mathcal{A}})$ is a one-by-one matrix whose trace just equals itself.

An $(n \times n)$-matrix whose entries lie in $\mathcal{L}(\hat{\mathcal{A}})$ gives a linear function $F \to F$ over each point in $[R/G]$. We can also restrict it to

$$f_{[w,e]} : F_{\mathrm{in}} \to F_{\mathrm{out}}$$

by taking an $(n_{\mathrm{out}} \times n_{\mathrm{in}})$-matrix whose entries γ_{jk} belong to $\mathfrak{e}_{\mathrm{out},k}^* \cdot \hat{\mathcal{A}} \cdot \mathfrak{e}_{\mathrm{in},j}$ where $(\mathfrak{e}_{\mathrm{in},j} : j = 1, \ldots, n_{\mathrm{in}})$ denotes the part of $(\mathfrak{e}_j : j = 1, \ldots, n)$ that has source in F_{in} (and similar for $\mathfrak{e}_{\mathrm{out},k}$). This produces a linear machine function $f^{\gamma}_{[w,e]}$ corresponding to a linear algorithm γ.

Example 3.12. *We can carry out linear machine learning for Example 3.1 by simply dropping the non-linear activation functions σ_i^1. Then the machine function is well-defined on the moduli space $[R/G]$ and we have a dimension reduction from R to $[R/G]$.*

However, keeping the non-linear activation functions is crucial for deep learning. We will make a non-linear algorithm well-defined on $[R/G]$ in the following subsections.

The cost function can also be defined algebraically as an element in $DR^0(\hat{\mathcal{A}}/B)$. Namely, given a function $f : F_{\mathrm{in}} \to F_{\mathrm{out}}$ and fixing $v \in F_{\mathrm{in}} = \mathbb{C}^{n_{\mathrm{in}}}$, the expression

$$\begin{aligned} E &= \int_K \left| \left(\sum_j \gamma_{jk}\, v_j : k = 1, \ldots, n_{\mathrm{out}} \right) - f\,(v) \right|^2_{F_{\mathrm{out}}} dv \\ &= \int_K \sum_k \left(\sum_j \gamma_{jk}\, v_j - f_k(v) \right) \left(\sum_j \gamma_{jk}^* \, \overline{v_j} - \overline{f_k(v)} \right) dv \end{aligned}$$

lies in $DR^0(\hat{\mathcal{A}}/B)$. Its differential in $DR^1(\hat{\mathcal{A}}/B)$ induces a one-form on $[R/G]$, which plays a central role in machine learning.

Suppose $\mathcal{A}$ is finitely generated, and so is $\hat{\mathcal{A}}$. Let $\{x_j : j = 1, \ldots, M\}$ be the generators of $\hat{\mathcal{A}}$. Then the algebraic Jacobian ring

$$DR^0(\hat{\mathcal{A}}/B)/\langle \partial_{x_j} E : j = 1, \ldots, M \rangle,$$

where $\partial_{x_j} E$ is the cyclic differential, is useful in capturing the critical locus of E.

3.4. Differential forms for near-ring

The associative algebra A in the last subsection captures linear operations of a computing machine, and has rich noncommutative geometries. In this subsection, we incorporate non-linear operations and extend the geometric construction to a near-ring.

3.4.1. Near-rings and their representations

Definition 3.13. *A near-ring is a set $\tilde{A}$ with two binary operations $+, \circ$ called addition and multiplication such that*

1. *$\tilde{A}$ is a group under addition.*
2. *Multiplication is associative.*
3. *Right multiplication is distributive over addition:*

$$(x + y) \circ z = x \circ z + y \circ z$$

for all $x, y, z \in \tilde{A}$.

In this paper, the near-ring we use will be required to satisfy that:

(4) $(\tilde{A}, +)$ is a vector space over $\mathbb{F} = \mathbb{C}$, with $c \cdot (x \circ y) = (c \cdot x) \circ y$ for all $c \in \mathbb{C}$ and $x, y \in \tilde{A}$.
(5) There exists $1 \in \tilde{A}$ such that $1 \circ x = x = x \circ 1$.

We call it a near-ring over $\mathbb{C}$ with identity, or a $\mathbb{C}$-near-ring with identity.

Note that $x \circ (c \cdot y) \neq c \cdot x \circ y$ in general. The following gives a prototype example.

Example 3.14. *The set* $\mathrm{Map}(V, V)$ *of $\mathbb{C}$-valued smooth functions $f : V \to V$ on a vector space V forms a near-ring over $\mathbb{C}$ with identity, with $+$ being the addition on the vector space, $\circ$ being the composition of functions, and* 1 *being the identity function on V.*

Definition 3.15. *Given a $\mathbb{C}$-near-ring with identity $\tilde{A}$, a $\mathbb{C}$-sub-near-ring is a $\mathbb{C}$-subspace $\tilde{A}' \subset \tilde{A}$ which is closed under the multiplication $\circ$. $\tilde{A}'$ is called a $\mathbb{C}$-sub-near-ring with identity if in addition, $1 \in \tilde{A}'$.*

Given an algebra A and a set S, we have the $\mathbb{C}$-near-ring $A\{S\}$ defined as follows.

Definition 3.16. *Let A be a $\mathbb{C}$-algebra with identity and S be a set. we define the $\mathbb{C}$-near-ring with identity $A\{S\}$ as follows. As a vector space,*

$$A\{S\} := \bigoplus_{p=0}^{\infty} A\{S\}_p$$

where:

1. *$A\{S\}_0 = A$;*
2. *Given $A\{S\}_p$ defined, $A\{S\}_{p+1}$ is spanned by the elements $a\varsigma \circ \alpha$, where $a \in A$, $\varsigma \in S$, and $\alpha \in A\{S\}_p$, subject to the relation $(a_1\varsigma_1 + ca_2\varsigma_2)\circ\alpha = a_1\varsigma_1 \circ \alpha + ca_2\varsigma_2 \circ \alpha$ for all $c \in \mathbb{C}$, $a_1, a_2 \in A$.*

Moreover, we define $1_A \circ \varsigma = \varsigma \circ 1_A = \varsigma$. Thus 1_A is also the identity for $A\{S\}$.

In the application to neural network, the elements $\varsigma \in S$ are symbols for the activation functions. Each element of $A\{S\}$ can be recorded by a rooted tree (oriented towards the root) defined as follows.

Definition 3.17. *Given $\tilde{A} = A\{S\}$, an activation tree is a rooted tree with the following labels.*

1. *Leaves and the root are labeled by $1_{\tilde{A}}$;*
2. *Edges are labeled by $a \in A$;*
3. *Nodes that are neither leaves nor the root are labeled by $\varsigma \in S$.*

Each node gives the output

$$\sum_k a_k \varsigma_k \circ \alpha_k, \tag{13}$$

where a_k are the labels of the incoming edges, ς_k and α_k are the labels of the tails of the incoming edges and their outputs respectively. (At a leaf, the label is $1_{\tilde{A}}$ and the output is $1_{\tilde{A}}$.) The element in $A\{S\}$ corresponding to the tree is the output of its root.

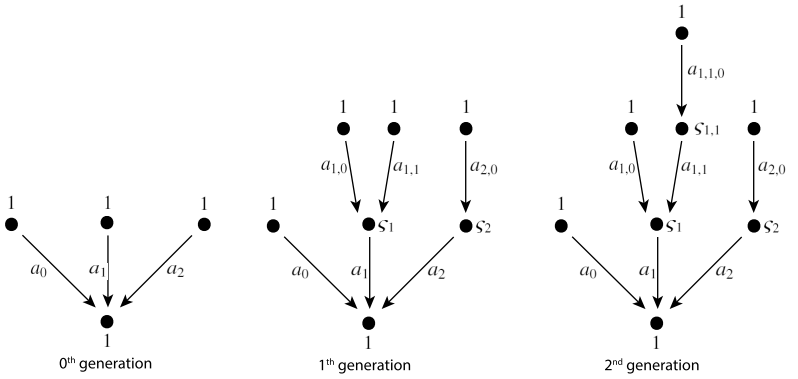

Figure 4

Remark 3.18. *The expression* (13) *takes the pre-activation value as the output of a node. One can also slightly modify the definition of an activation tree and use the other convention that takes the activation value as the output.*

Example 3.19. *Figure 4 shows examples of activation trees that represent elements in* $A\{S\}$. *The expression corresponding to the rightmost tree is*

$$a_0 + a_1\varsigma_1 \circ (a_{1,0} + a_{1,1}\varsigma_{1,1} \circ a_{1,1,0}) + a_2\varsigma_2 \circ a_{2,0}$$

for some $a_0, a_1, a_{1,0}, a_{1,1}, a_{1,1,0}, a_{2,0} \in A$, $\varsigma_1, \varsigma_{1,1}, \varsigma_2 \in S$.

Note that the tree here is not the digraph (quiver) that we will consider in the later part of this paper. The labels a for the edges will be taken to be elements in the double of a quiver algebra $\hat{\mathcal{A}}$ *later, and required to be loops from the framing of the quiver back to itself.*

The above definition goes from a $\mathbb{C}$-algebra to a $\mathbb{C}$-near-ring. In the reverse direction, we can define the following.

Definition 3.20. *The canonical subalgebra of a* $\mathbb{C}$*-near-ring* $\tilde{A}$ *with identity is defined as*

$$A := \{x \in \tilde{A} : x \circ (cy + z) = cx \circ y + x \circ z \text{ for all } y, z \in \tilde{A} \text{ and } c \in \mathbb{C}\}.$$

It is easy to check that

Lemma 3.21. *A is a* $\mathbb{C}$*-algebra with identity.*

Example 3.22. *For the above example that* $\tilde{A} = \mathrm{Map}(V, V)$, *the canonical subalgebra is the subset* $\mathrm{End}(V)$ *of linear endomorphisms of* V. *This can be*

seen by taking $y, z \in \mathrm{Map}(V, V)$ to be constant maps in the above definition of A.

Given a subset S of $\tilde{A}$, we have the sub-near-ring generated by S defined as follows.

Definition 3.23. *The sub-near-ring of $\tilde{A}$ generated by S, which is denoted as $\langle S \rangle_{\tilde{A}}$, is defined inductively as follows. As a vector space,*

$$\langle S \rangle_{\tilde{A}} := \sum_{p=0}^{\infty} \langle S \rangle_{\tilde{A},p} \subset \tilde{A}$$

where:

1. $\langle S \rangle_{\tilde{A},0} = A$;
2. *Given $\langle S \rangle_{\tilde{A},p}$ defined, $\langle S \rangle_{\tilde{A},p+1}$ is spanned by the elements $a \circ \varsigma \circ \alpha$, where $a \in A$, $\varsigma \in S$, and $\alpha \in \langle S \rangle_{\tilde{A},p}$.*

$\tilde{A}$ is said to be finitely generated if $\tilde{A} = \langle S \rangle_{\tilde{A}}$ for a finite subset $S \subset \tilde{A}$. $S \subset \tilde{A}$ is said to be a free generating subset if $\langle S \rangle_{\tilde{A}} = A\{S\}$.

It is easy to check that:

Proposition 3.24. *$\langle S \rangle_{\tilde{A}}$ defined above is a sub-near-ring.*

Example 3.25. *Let's continue the example of the set of functions $\mathrm{Map}(V, V)$. Fix a collection of non-linear functions $\sigma_1, \ldots, \sigma_N : V \to V$. This corresponds to a finitely generated sub-near-ring $A\{\sigma_1, \ldots, \sigma_N\} \subset \tilde{A}$. $\sigma_1, \ldots, \sigma_N$ can be chosen such that they are not related by iterated compositions and linear combinations. Then they form a free generating subset.*

Definition 3.26. *A morphism of $\mathbb{C}$-near-rings with identities is a map $\Psi : \tilde{A}_1 \to \tilde{A}_2$ that satisfies:*

1. $\Psi(x + y) = \Psi(x) + \Psi(y)$;
2. $\Psi(x \circ y) = \Psi(x) \circ \Psi(y)$;
3. $\Psi(1_{\tilde{A}_1}) = 1_{\tilde{A}_2}$.

Ψ is said to be a strong morphism if in addition, it satisfies:

(4) Ψ maps the canonical subalgebra of $\tilde{A}_1$ to that of $\tilde{A}_2$.

It easily follows from the definition that a surjective morphism of $\mathbb{C}$-near-rings is automatically strong.

Now we consider modules of a $\mathbb{C}$-near ring.

Definition 3.27. *For a $\mathbb{C}$-near ring $\tilde{A}$ with identity, an $\tilde{A}$-module is a $\mathbb{C}$-vector space V together with a strong $\mathbb{C}$-near-ring morphism $\tilde{A} \to \mathrm{Map}(V, V)$.*

For two $\tilde{A}$-modules V, W, a morphism from V to W is a map $\phi \in \mathrm{Map}(V, W)$ that commutes with the actions of $\tilde{A}$:

$$\phi \circ V(\alpha)(v) = W(\alpha) \circ \phi(v)$$

for all $\alpha \in \tilde{A}$.

It follows from the above definition that an $\tilde{A}$-module is automatically an A-module (where A denotes the canonical subalgebra).

Essentially, the method of deep learning is performing a (stochastic) gradient descent on a certain subvariety of the space of $\tilde{A}$-modules for a fixed near-ring $\tilde{A}$. However, such a space of $\tilde{A}$-modules is typically infinite-dimensional (since the choice of non-linear maps is infinite-dimensional). We would like to systematically construct explicit $\tilde{A}$-modules. A useful construction for $\tilde{A} = A\{\varsigma_1, \dots, \varsigma_N\}$ is the following. Given an algebra and an A-module V, a choice of $\sigma_1, \dots, \sigma_N \in \mathrm{Map}(V, V)$ enhances V to be an $A\{\varsigma_1, \dots, \varsigma_N\}$-module. (Here, ς_l are the formal symbols corresponding to σ_l.)

Unfortunately, such a correspondence between A-modules and $\tilde{A}$-modules does not behave well in the morphism level. Namely, an A-module endomorphism $\phi \in \mathrm{Hom}(V, V)$ typically does not satisfy $\phi \circ \sigma_l = \sigma_l \circ \phi$ for non-linear functions $\sigma_l \in \mathrm{Map}(V, V)$, and hence cannot be lifted as an $\tilde{A}$-module morphism. So we do not have a map from the space of A-modules to the space of $\tilde{A}$-modules that descends to isomorphism classes.

Below, we use our setting of an activation module to remedy this correspondence between A and $\tilde{A}$. See Proposition 3.29.

3.4.2. Forms over near-ring Let $\mathcal{A}$ be an algebra, and fix a framing vector space $F = \mathbb{C}^n$. In Section 3.3, we have taken the doubled augmented algebra $\hat{\mathcal{A}}$. Now, we consider the set $\mathrm{Mat}_F(\hat{\mathcal{A}})$ of $n \times n$ matrices whose (k, j)-th entries lie in $\mathfrak{e}_k^* \cdot \hat{\mathcal{A}} \cdot \mathfrak{e}_j$.

It is easy to check that:

Lemma 3.28. *$A := \mathrm{Mat}_F(\hat{\mathcal{A}})$ forms an algebra under matrix addition and multiplication (where multiplication between entries is given by $\hat{\mathcal{A}}$).*

This is essentially the algebra $\mathcal{L}(\mathcal{A})$ defined in Equation (7), adapted to the current setting by identifying $\mathfrak{e} = (\mathfrak{e}_j : j = 1, \dots, n)$. As explained previously right after Corollary 3.11, each element of $\mathrm{Mat}_F(\hat{\mathcal{A}})$ induces a section of the trivial bundle $\mathrm{End}(F)$ over $[R/G]$, where R is the space of framed representations of $\mathcal{A}$.

Similar to (8), we take the $\mathbb{C}$-near-ring

$$\widetilde{\mathcal{A}} := \mathrm{Mat}_F(\hat{\mathcal{A}})\{\varsigma_1, \ldots, \varsigma_N\}$$

where each ς_l represents a non-linear function $\sigma_l : F \to F$.

As in Definition 3.16, we have a natural grading on $\widetilde{\mathcal{A}}$. Recall that the elements of $\widetilde{\mathcal{A}}$ can be recorded by rooted trees. The generation of rooted trees gives a grading on $\widetilde{\mathcal{A}}$:

$$\widetilde{\mathcal{A}} = \bigoplus_k \widetilde{\mathcal{A}}_k.$$

$\widetilde{\mathcal{A}}_0 = \mathrm{Mat}_F(\hat{\mathcal{A}})$; $\widetilde{\mathcal{A}}_p$ consists of linear combinations of $a \cdot \varsigma_j \circ \alpha$ for $a \in \mathrm{Mat}_F(\hat{\mathcal{A}})$, $\alpha \in \widetilde{\mathcal{A}}_{p-1}$, and $j = 1, \ldots, N$.

In the last subsection, we have explained a correspondence between A-modules and $A\{\varsigma_1, \ldots, \varsigma_N\}$-modules, by choosing maps $\sigma_1, \ldots, \sigma_N$

$\in \mathrm{Map}(V, V)$. However, such a correspondence does not descend to isomorphism classes. The advantage of the construction here (after fixing a framing vector space F) is that the correspondence is well-defined on the moduli space.

Proposition 3.29. *Fix $\sigma_l^F \in \mathrm{Map}(F, F)$ for $l = 1, \ldots, N$. A framed $\mathcal{A}$-module (V, w, e) with a Hermitian metric h on V induces an $\widetilde{\mathcal{A}}$-module structure on F. Moreover, if two such modules with metrics are isomorphic*

$(V, w, e, h) \cong (V', w', e', h')$, then the induced $\widetilde{\mathcal{A}}$-module structures on F are the same. Thus, fixing an equivariant family of metrics on V, we have the map

$$[R(A)/G] \to R_F\left(\widetilde{\mathcal{A}}\right)$$

where $R_F(\widetilde{\mathcal{A}})$ denotes the space of $\widetilde{\mathcal{A}}$-module structures on F.

Proof. As explained below Corollary 3.11, by using the framed $\mathcal{A}$-module structure and metric, each element in $\mathrm{Mat}_F(\hat{\mathcal{A}})$ induces a linear endomorphism of F, which is invariant under $\mathrm{GL}(V)$. Thus two isomorphic framed modules with metrics produce the same linear endomorphism of F. Moreover, σ_l^F are maps on F which receive no action by $\mathrm{GL}(V)$. As a result, this gives an $\widetilde{A}$-module structure on F which remains the same for isomorphic (V, w, e, h). □

The above proposition explains why we want Definition 3.5 for an activation module.

Remark 3.30. *In Definition 3.5, we have a splitting $F = F_{\mathrm{m}} \oplus F_{\mathrm{in}} \oplus F_{\mathrm{out}}$. It is easy to restrict to the component F_{m} (or other components). We have the projection $p : F \to F_{\mathrm{m}}$ and inclusion $\iota : F_{\mathrm{m}} \to F$. The functions $\sigma_j^F : F_{\mathrm{m}} \to F_{\mathrm{m}}$ can also be understood as functions on F. From now on, we will simply work with the whole framing vector space F, keeping in mind that we can restrict to the components if we want.*

We are going to define differential forms on $\widetilde{\mathcal{A}}$. Under the setting of Definition 3.5, they will induce $\mathrm{Map}\,(F, F)$-valued forms on $[R/G]$ (Theorem 3.42).

First, recall that we have the Karoubi-de Rham complex $DR^\bullet(\hat{\mathcal{A}}/B)$. It contains the subspace of forms over loops at the framing vertex. These forms are linear combinations of elements

$$\mathfrak{e}_k^* \dots \mathfrak{e}_j, (d\mathfrak{e}_k^*) \dots \mathfrak{e}_j, \mathfrak{e}_k^* \dots (d\mathfrak{e}_j), (d\mathfrak{e}_k^*) \dots (d\mathfrak{e}_j)$$

for some $j, k = 1, \dots, n$. In other words, the subspace is $\sum_{j,k=1}^n DR^\bullet(\hat{\mathcal{A}}/B)_{j,k}$, where $DR^\bullet(\hat{\mathcal{A}}/B)_{j,k}$ is defined as

$$\mathfrak{e}_k^* \cdot DR^\bullet(\hat{\mathcal{A}}/B) \cdot \mathfrak{e}_j + d\mathfrak{e}_k^* \cdot DR^\bullet(\hat{\mathcal{A}}/B) \cdot \mathfrak{e}_j + \mathfrak{e}_k^* \cdot DR^\bullet(\hat{\mathcal{A}}/B) \cdot d\mathfrak{e}_j + d\mathfrak{e}_k^* \cdot DR^\bullet(\hat{\mathcal{A}}/B) \cdot d\mathfrak{e}_j.$$

We define the linear part as follows.

Definition 3.31. *$DR^\bullet(\mathrm{Mat}_F(\hat{\mathcal{A}}))$ is defined to be the space of $n \times n$ matrices whose (k, j)-th entries lie in $DR^\bullet(\hat{\mathcal{A}}/B)_{j,k}$.*

Like $DR^\bullet(\hat{\mathcal{A}}/B)$, this space is graded by the degree of forms.

From Proposition 3.9, we have the map

$$DR^\bullet(\mathrm{Mat}_F(\hat{\mathcal{A}})) \to (\Omega^\bullet(R, \mathrm{End}\,(F)))^G. \tag{14}$$

(F, and hence $\mathrm{End}(F)$, are treated as a trivial bundle over $[R/G]$.)

To define differential forms on $\widetilde{\mathcal{A}}$, we need to use the symbols $D^{(p)}\varsigma_l\big|_\alpha\,(a_1, \dots, a_p)$, which represent the p-th order *symmetric* differentials of the non-linear functions σ_l. For instance, $D^{(1)}\varsigma_l$ represents the usual differential $d\sigma_l$; $D^{(2)}\varsigma_l$ represents the Hessian of σ_l, which is a symmetric bilinear two-form. $D^{(p)}\varsigma_l$ is supersymmetric about its p inputs:

$$\begin{aligned} &D^{(p)}\varsigma_l\,|_\alpha(a_1, \dots, a_k, a_{k+1}, \dots, a_p) \\ &= (-1)^{\deg a_k \cdot \deg a_{k+1}} D^{(p)}\varsigma_l\,|_\alpha(a_1, \dots, a_{k+1}, a_k, \dots, a_p) \end{aligned} \tag{15}$$

where $\deg a$ denotes the degree of a. The inputs a_i are again differential forms on $\widetilde{\mathcal{A}}$. The point of evaluation α is an element of $\widetilde{\mathcal{A}}$.

Definition 3.32. *A form-valued tree is a rooted tree (oriented towards the root) whose edges are labeled by* $\phi \in DR^\bullet(\mathrm{Mat}_F(\hat{\mathcal{A}}))$*; leaves are labeled by* $\alpha \in \widetilde{\mathcal{A}}$*; the root (if not being a leaf) is labeled by* 1*; nodes which are neither leaves nor the root are labeled by* $D^{(p)}\varsigma_l\Big|_\alpha$ *for some* $l = 1, \ldots, N$*,* $\alpha \in \widetilde{\mathcal{A}}$*, and* $p > 0$ *is the number of incoming edges.*

The trivial rooted tree, which has a single node with no edge, corresponds to a zero-form. The node is attached with an element $\alpha \in \widetilde{\mathcal{A}}$.

For a non-trivial rooted tree, the output of each node which is neither a leaf nor the root is

$$D^{(p)}\varsigma_l\Big|_\alpha (\phi_1 \cdot \eta_1, \ldots, \phi_p \cdot \eta_p)$$

where $\phi_k \in DR^\bullet(\mathrm{Mat}_F(\hat{\mathcal{A}}))$ are attached to the incoming edges, and η_k are the outputs of the nodes adjacent to the incoming edges. The input edges to the node are read clockwisely. Its degree is defined as the sum of $\deg(\phi_k \cdot \eta_k) = \deg \phi_k + \deg \eta_k$. The output of each leaf is simply its label $\alpha \in \widetilde{\mathcal{A}}$ which has degree 0. The output of the root, which is the sum of $\phi_k \cdot \eta_k$ for the incoming edges ϕ_k and outputs of incoming nodes η_k, is taken to be the differential form associated to the form-valued tree.

Remark 3.33. *Now we have introduced two different kinds of rooted trees. The activation tree represents an element in* $\widetilde{\mathcal{A}}$ *(which is identified as a zero-form); the form-valued tree represents a p-form. For* $p = 0$*, the form-valued tree is trivial consisting of a single root, which is labeled by* $\alpha \in \widetilde{\mathcal{A}}$*.* α *is represented by an activation tree, which is more useful in this situation.*

Definition 3.34. *A differential zero-form over* $\widetilde{\mathcal{A}}$ *is simply an element in* $\widetilde{\mathcal{A}}$*. Denote*

$$DR^0(\widetilde{\mathcal{A}}) := \widetilde{\mathcal{A}}.$$

A differential p-form (for $p \geq 1$*) is a sum of forms associated to form-valued trees with at most p leaves, with total of degrees of forms attached to edges being p. The space of p-forms is denoted by* $DR^p(\widetilde{\mathcal{A}})$*.*

Remark 3.35. *Since we require the trees contributing to a p-form to have at most p leaves,* $D^{(k)}\varsigma_l$ *that appear at the nodes must have* $k \leq p$*.*

Example 3.36. *Figure 5 shows examples of one-forms and two-forms. They correspond to* $a_1 da_2 \cdot D^{(1)}\varsigma_l|_{\alpha_1}(a_3 \cdot \alpha_2)$*,* $a_1 da_2 \cdot D^{(1)}\varsigma_l|_{\alpha_1}((a_3 da_4) \cdot \alpha_2)$ *and* $a_1 da_2 \cdot D^{(2)}\varsigma_l|_{\alpha_1}(a_3 \cdot \alpha_2, (a_4 da_5) \cdot \alpha_3)$ *respectively.*

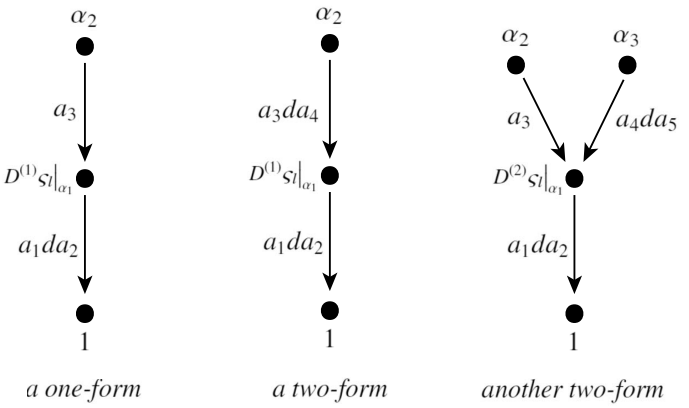

Figure 5

Definition 3.37. *The differential of a form over $\widetilde{\mathcal{A}}$ is defined as follows.*

A zero-form in the 0-th graded piece $\alpha \in \widetilde{\mathcal{A}}_0$ is simply an element in $\mathrm{Mat}_F(\hat{\mathcal{A}})$, and its differential is given by entrywise differential in $DR^\bullet(\hat{\mathcal{A}}/B)$. A zero-form in the p-th graded piece $\alpha \in \widetilde{\mathcal{A}}_p$ can be written as

$$\alpha = a_0 + \sum_{k=1}^{m} a_k \circ \varsigma_{l(k)} \circ \alpha_k \in DR^0(\widetilde{\mathcal{A}})$$

where $a_k \in \mathrm{Mat}_F(\hat{\mathcal{A}})$ for $k = 0, \ldots, m$, $\alpha_k \in \widetilde{\mathcal{A}}_{p-1}$, and $l(k) = 1, \ldots, N$. Then

$$d\alpha := da_0 + \sum_k da_k \cdot (\varsigma_{l(k)} \circ \alpha_k) + \sum_k a_k \cdot D^{(1)}\varsigma_{l(k)}\Big|_{\alpha_k}(d\alpha_k) \in DR^1(\widetilde{\mathcal{A}})$$

where $d\alpha_k$ has already been defined by the inductive assumption since $\alpha_k \in \widetilde{\mathcal{A}}_{p-1}$.

For p-forms with $p > 0$, it suffices to define differential of a p-form attached to a form-valued tree. For a leaf, the output is simply its label $\alpha \in \widetilde{\mathcal{A}}$, whose differential has been defined above. For a node which is neither a leaf nor the root, its output is of the form $D^{(p)}\varsigma_l\Big|_\alpha(\phi_1 \cdot \eta_1, \ldots, \phi_p \cdot \eta_p)$, where $\phi_k \in DR^\bullet(\mathrm{Mat}_F(\hat{\mathcal{A}}))$ are attached to the incoming edges, and η_k are the outputs of the nodes adjacent to the incoming edges. Its differential is defined as

$$d\left(D^{(p)}\varsigma_l\Big|_\alpha(\phi_1 \cdot \eta_1, \ldots, \phi_p \cdot \eta_p)\right)$$

$$
\begin{aligned}
&:= D^{(p+1)}\varsigma_l\Big|_\alpha (d\alpha, \phi_1 \cdot \eta_1, \dots, \phi_p \cdot \eta_p) \\
&\quad + \sum_{k=1}^{p} (-1)^{\deg(\phi_1\eta_1)+\dots+\deg(\phi_{k-1}\eta_{k-1})}\, D^{(p)}\varsigma_l\Big|_\alpha (\phi_1 \cdot \eta_1, \dots, (d\phi_k) \cdot \eta_k \\
&\quad + (-1)^{\deg \phi_k} \phi_k \cdot d\eta_k, \dots, \phi_p \cdot \eta_p)
\end{aligned}
$$

where the differential $d\eta_k$ is already known by induction assumption on the generation of the tree. The p-form attached to the tree is the output of the root, which is of the form $\sum_k \phi_k \cdot \eta_k$. Its differential is defined as

$$\sum_k \left(d\phi_k \cdot \eta_k + (-1)^{\deg \phi_k} \phi_k \cdot d\eta_k \right)$$

where $d\eta_k$ has been defined by inductive assumption.

The differential of a zero-form has a nice expression in terms of a sum over sub-trees of the activation tree as follows.

Proposition 3.38. *Consider $\alpha \in \widetilde{\mathcal{A}}$ represented by an activation tree T. Then $d\alpha \in DR^1(\widetilde{\mathcal{A}})$ is a sum over all the nodes of T, and the terms are given as follows. For each node, there is a unique path $\gamma_1 \dots \gamma_r$ in T connecting from that node to the root, where γ_k denotes the (oriented) edges. (When the node is the root, the path is trivial and the corresponding term is simply 0.) The corresponding term is equal to*

$$a_{\gamma_1} D^{(1)}\varsigma_{l(t(\gamma_1))}\Big|_{\alpha_{t(\gamma_1)}} \dots a_{\gamma_{r-1}} D^{(1)}\varsigma_{l(t(\gamma_{r-1}))}\Big|_{\alpha_{t(\gamma_{r-1})}} da_{\gamma_r} \cdot (\varsigma_{l(t(\gamma_r))} \circ \alpha_{t(\gamma_r)}) \tag{16}$$

where α_i for a node i of T denotes the output at the node i.

Proof. The statement easily holds for the zeroth generation: the tree only has the root and leaves as nodes, and the zeroth form has an expression $\sum_i a_i$ for $a_i \in DR^\bullet(\mathrm{Mat}_F(\hat{\mathcal{A}}))$, whose differential is simply $\sum_i da_i$, which is a sum over the leaves.

Suppose the statement holds for all elements in the p-th generation. For $\alpha = a_0 + \sum_{k=1}^m a_k \circ \varsigma_{l(k)} \circ \alpha_k$ in the $(p+1)$-th generation, $d\alpha = da_0 + \sum_k da_k \cdot (\varsigma_{l(k)} \circ \alpha_k) + \sum_k a_k \cdot D^{(1)}\varsigma_{l(k)}\Big|_{\alpha_k} (d\alpha_k) \in DR^1(\widetilde{\mathcal{A}})$, where $d\alpha_k$ is a sum over the nodes of the activation tree of α_k as given in Equation (16). The first term da_0 and second term $da_k \cdot (\varsigma_{l(k)} \circ \alpha_k)$ correspond to the tail nodes of the edges of a_k for $k = 0, \dots, m$. Thus $d\alpha$ is a sum over all the nodes with the summands given by (16). □

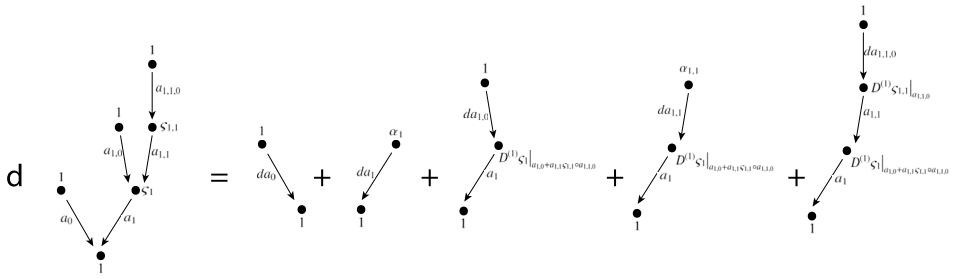

Figure 6

Example 3.39. *Consider the* 0*-form*

$$\alpha = a_0 + a_1 \varsigma_1 \circ (a_{1,0} + a_{1,1} \varsigma_{1,1} \circ a_{1,1,0}).$$

Its differential is equal to

$$\begin{aligned} d\alpha &= da_0 + da_1 \cdot \alpha_1 \\ &\quad + a_1 \, D^{(1)}\varsigma_1\Big|_{a_{1,0}+a_{1,1}\varsigma_{1,1}\circ a_{1,1,0}} (da_{1,0} + da_{1,1} \cdot \alpha_{1,1} + a_{1,1} \, D^{(1)}\varsigma_{1,1}\Big|_{a_{1,1,0}} da_{1,1,0}) \end{aligned}$$

where $\alpha_1 = \varsigma_1 \circ (a_{1,0} + a_{1,1}\varsigma_{1,1} \circ a_{1,1,0})$ *and* $\alpha_{1,1} = \varsigma_{1,1} \circ a_{1,1,0}$*. It is equal to the sum over the nodes of the activation tree of* α *as shown in Figure 6.*

Remark 3.40. *The output at a node of an activation tree representing* $\alpha \in \widetilde{\mathcal{A}}$ *can be computed by the algorithm called* forward propagation*. Namely, the previous results* α_k *(pre-activation values) have been stored in memory, and the current output is computed as* $\sum_k a_k \cdot \varsigma_{l(k)} \circ \alpha_k$ *(where* $\varsigma_{l(k)}$ *are the activation functions at previous nodes and* a_k *are labeling the incoming edges) and stored to memory for later steps.*

For the differential $d\alpha$*, the computation* (16) *uses the stored outputs* α_i *in the forward propagation. Moreover, the expression*

$$a_{\gamma_1} \, D^{(1)}\varsigma_{l(t(\gamma_1))}\Big|_{\alpha_{t(\gamma_1)}} \ldots a_{\gamma_{r-1}} \, D^{(1)}\varsigma_{l(t(\gamma_{r-1}))}\Big|_{\alpha_{t(\gamma_{r-1})}}$$

appears in every term of $d\alpha$ *corresponding to a path in* T *that contains* $\gamma_{r-1} \ldots \gamma_1$*. Thus it is good to start with the root to compute and store the values of* $a_{\gamma_1} \, D^{(1)}\varsigma_{l(t(\gamma_1))}\Big|_{\alpha_{t(\gamma_1)}} \ldots a_{\gamma_{r-1}} \, D^{(1)}\varsigma_{l(t(\gamma_{r-1}))}\Big|_{\alpha_{t(\gamma_{r-1})}}$*, and move backward with respect to the orientation of the tree* T*. This is well known as the* backward propagation algorithm.

Proposition 3.41. $d^2 = 0$.

Proof. First consider a zero-form, that is, $\alpha \in \widetilde{\mathcal{A}}$. α is represented by an activation tree. Recall that $d^2 a = 0$ for $a \in A = \mathrm{Mat}_F(\hat{\mathcal{A}})$ as these elements are zero-forms defined over the associative algebra A in the standard Karoubi-de Rham sense.

We can write

$$\alpha = a_0 + \sum_{k=1}^{m} a_k \circ \varsigma_{l(k)} \circ \alpha_k \in DR^0(\widetilde{\mathcal{A}})$$

where $a_k \in \mathrm{Mat}_F(\hat{\mathcal{A}})$ for $k = 0, \ldots, m$, $\alpha_k \in \widetilde{\mathcal{A}}$ has one less generation than α, and $l(k) = 1, \ldots, N$. Then

$$\begin{aligned} d^2\alpha =& d\left(da_0 + da_k \cdot (\varsigma_{l(k)} \circ \alpha_k) + a_k \cdot D^{(1)}\varsigma_{l(k)}\Big|_{\alpha_k}(d\alpha_k) \right) \\ =& - da_k \cdot d(\varsigma_{l(k)} \circ \alpha_k) + da_k \cdot D^{(1)}\varsigma_{l(k)}\Big|_{\alpha_k}(d\alpha_k) \\ &+ a_k \cdot D^{(2)}\varsigma_{l(k)}\Big|_{\alpha_k}(d\alpha_k, d\alpha_k). \end{aligned}$$

The first two terms cancel since $d(\varsigma_{l(k)} \circ \alpha_k) = D^{(1)}\varsigma_{l(k)}\Big|_{\alpha_k}(d\alpha_k)$. The third term vanishes since $D^{(2)}\varsigma_{l(k)}$ is supersymmetric about its input (Equation (15)).

For a general p-form, it suffices to prove $d\psi = 0$ for ψ represented by a form-valued tree. We will do induction on the generation of the tree. We already know the statement when the tree is trivial (which is the case of a zero-form). The p-form ψ is given as $\psi = \sum_k \phi_k \cdot \eta_k$ for some $\phi_k \in DR^\bullet(\mathrm{Mat}_F(\hat{\mathcal{A}}))$ and η_k has a smaller generation than ψ. Then

$$\begin{aligned} d^2\psi =& \sum_k \Big((-1)^{\deg \phi_k} d\phi_k \cdot d\eta_k + (-1)^{\deg \phi_k + 1} d\phi_k \cdot d\eta_k \\ &+ (-1)^{2\deg \phi_k} \phi_k \cdot d^2\eta_k \Big). \end{aligned}$$

The first two terms cancel. The last term vanishes by inductive assumption. □

Finally, we show that differential forms on the near-ring $\widetilde{\mathcal{A}}$ induce G-invariant $\mathrm{Map}\,(F, F)$-valued differential forms over the space of framed $\mathcal{A}$-modules R.

Theorem 3.42. *There exists a degree-preserving map*

$$DR^\bullet(\widetilde{\mathcal{A}}) \to (\Omega^\bullet(R, \mathbf{Map}\,(F, F)))^G$$

which commutes with d on the two sides, and is equal to the map (14)*:* $DR^\bullet(\mathrm{Mat}_F(\hat{\mathcal{A}})) \to (\Omega^\bullet(R, \mathrm{End}\,(F)))^G$ *when restricted to* $DR^\bullet(\mathrm{Mat}_F(\hat{\mathcal{A}}))$*. Here,* $\mathbf{Map}\,(F,F)$ *denotes the trivial bundle* $\mathrm{Map}\,(F,F) \times R$*, and the action of* $G = \mathrm{GL}(V)$ *on fiber direction is trivial.*

Proof. First consider the case of a zero-form. We associate $\alpha \in DR^0(\widetilde{\mathcal{A}})$ to a G-invariant $\mathrm{Map}(F,F)$-valued function over R inductively on its generation as an element in $\widetilde{\mathcal{A}}$. In the zeroth generation, it is just an element in $\mathrm{Mat}_F(\hat{\mathcal{A}})$, which induces a matrix whose entries lie in $\Omega^0(R)^G$ by Proposition 3.9. This gives a self-map $F \to F$ over $[R/G]$. If α is in the p-th generation, then it is written as $\alpha = a_0 + \sum_{k=1}^m a_k \circ \varsigma_{l(k)} \circ \alpha_k \in DR^0(\widetilde{\mathcal{A}})$, where α_k is in the $(p-1)$-th generation and induces a self-map $F \to F$ over $[R/G]$. By composing with the corresponding functions $\sigma_{l(k)} : F \to F$ and the induced functions of $a_k \in \mathrm{Mat}_F(\hat{\mathcal{A}})$, we obtain a self-map $F \to F$ over $[R/G]$ corresponding to α.

For a k-form $\psi \in DR^\bullet(\widetilde{\mathcal{A}})$, we do an induction on the generation of its corresponding form-valued tree to associate it with a G-invariant $\mathrm{Map}\,(F,F)$-valued k-form over R. In th zeroth generation it must be a zero-form (where the associated form-valued tree is simply a single node), which is done by the previous paragraph. In general $\psi = \sum_k \phi_k \cdot \eta_k$ for some $\phi_k \in DR^\bullet(\mathrm{Mat}_F(\hat{\mathcal{A}}))$ and η_k has a smaller generation than ψ. Both ϕ_k and η_k have been associated with G-invariant $\mathrm{Map}\,(F,F)$-valued k-forms. Then their matrix products (and by wedge product entrywise) give the required k-form associated to ψ.

It follows from the chain rule that the differential for $DR^\bullet(\widetilde{\mathcal{A}})$ given in Definition 3.37 agrees with that for $\mathrm{Map}\,(F,F)$-valued forms over R. Moreover, for $\phi \in DR^\bullet(\mathrm{Mat}_F(\hat{\mathcal{A}}))$, it is in the first generation written as $\phi \cdot 1$. By the above definition, the association is given by the map (14). □

So far, this gives matrix-valued differential forms on $[R/G]$. To produce $\mathbb{C}$-valued forms, that is, to remove the component $\mathrm{Map}\,(F,F)$ in the above theorem, we proceed as follows. The near-ring $\widetilde{\mathcal{A}}$ can be augmented with the inclusion and projection symbols ι_i and p_j, where ι_i represents the inclusion $\mathbb{C} \to F$ of the i-th coordinate axis, and p_j represents the projection $F \to \mathbb{C}$ in the i-th direction. This forms an augmented near-ring

$$\bigoplus_{k=1}^{\infty} \left(\{p_1, \ldots, p_n\} \circ \widetilde{\mathcal{A}} \circ (\mathbb{C} \cdot \{\iota_1, \ldots, \iota_n\}) \right)^k$$

consisting of linear combinations of elements $\left(p_i \circ \alpha \circ \left(\sum_j x_j \iota_j \right) \right)^k$ for $\alpha \in \widetilde{\mathcal{A}}$, with the relations $p_i \circ \iota_j = \delta_{ij} \cdot 1$ and $\iota_j \circ 1_{\widetilde{\mathcal{A}}} \circ p_i = \delta_{ij} \cdot 1$. Then differential forms in this augmented near-ring induces G-invariant differential forms in $(\Omega^\bullet(R))^G$. The proof is similar and we shall not repeat.

In application, we fix an algorithm $\tilde{\gamma} \in \widetilde{\mathcal{A}}$ and consider

$$\varphi^{\tilde{\gamma}}(x) = \left(p_i \circ \tilde{\gamma} \circ \left(\sum_j x_j \iota_j \right) \right)_{i=1}^{n}$$

for each element $x = (x_1, \ldots, x_n) \in F$. $\varphi^{\tilde{\gamma}}(x)$ is a vector whose entries are elements inside the above augmented near-ring. Given $f : K \to F$, we have

$$\int_K \left| \varphi^{\tilde{\gamma}}(x) - f(x) \right|^2 dx$$

which is a 0-form on the augmented near-ring. This 0-form and its differential induces the cost function and its differential on $[R/G]$ respectively, which are the central objects in machine learning.

Example 3.43. *Continuing Example 3.1, we now define an algorithm $\tilde{\gamma}$ over the moduli space $[R/G]$.*

$\mathfrak{e}_m$ consists of two parts $\mathfrak{e}_{m,1}$ and $\mathfrak{e}_{m,2}$ that correspond to the weights and biases. $\tilde{\gamma}$ is taken to be

$$\tilde{\gamma} = \hat{a}_2 \varsigma \circ \hat{a}_1$$

where $\hat{a}_1 = (\mathfrak{e}_{m,1})^*(a^1 \mathfrak{e}_{in} + \mathfrak{e}_{m,2})$ and $\hat{a}_2 = (\mathfrak{e}_{out})^* a^2 \mathfrak{e}_{m,1}$. Note that both $\hat{a}_1$ and $\hat{a}_2$ involve $\mathfrak{e}_m$. We have

$$\varphi^{\tilde{\gamma}}(x) = \left(p_k \circ \hat{a}_2 \varsigma \circ \hat{a}_1 \circ \left(\sum_j^{784} x_j \iota_j \right) \right)_{k=1}^{10}.$$

From the 0-form $\int_K \left| \varphi^{\tilde{\gamma}}(x) - f(x) \right|^2 dx$, we calculate the differential as

$$d\left(\int_K \left| \varphi^{\tilde{\gamma}}(x) - f(x) \right|^2 dx \right) = 2 \left(\varphi^{\tilde{\gamma}}(x) - f(x) \right) \cdot d\varphi^{\tilde{\gamma}}$$

where

$$\begin{aligned} (d\varphi^{\tilde{\gamma}}(x))_i = {} & dp_i \cdot \hat{a}_2 \varsigma \circ \hat{a}_1 \circ \left(\sum_j^{784} x_j \iota_j \right) + p_i (d\hat{a}_2) \varsigma \circ \hat{a}_1 \circ \left(\sum_j^{784} x_j \iota_j \right) \\ & + p_i \circ \hat{a}_2 D^{(1)} \varsigma_{\hat{a}_1 \circ \left(\sum_j^{784} x_j \iota_j \right)} \left(d\hat{a}_1 \cdot \left(\sum_j^{784} x_j \iota_j \right) + \hat{a}_1 \left(\sum_j^{784} x_j d\iota_j \right) \right). \end{aligned} \tag{17}$$

Over each point $[w_1, w_2, b]$ in $[R/G]$, $\varphi^{\tilde{\gamma}}(x)$ and the 1-form $d\varphi^{\tilde{\gamma}}(x)$ induce a machine function and its differential respectively.

4. Uniformization

In this section, we apply the idea of uniformization of metrics on framed quiver moduli spaces, which are interpreted as moduli of computing machines as in the previous section.

The uniformization theorem for Riemann surfaces was a big discovery of Klein, Poincaré and Koebe in the 19th century. It asserts that every simply connected Riemann surface is conformally equivalent to either the complex plane, the Riemann sphere, or the hyperbolic disc.

Such a classification also holds for Riemannian symmetric spaces. Namely, any irreducible simply connected symmetric space is either of Euclidean type, compact type, or non-compact type, depending on whether its sectional curvature is identically zero, non-negative, or non-positive.

As a key example, $\mathrm{Gr}(n,d)$ is a compact Hermitian symmetric space. It has a non-compact dual which embeds as an open subset of $\mathrm{Gr}(n,d)$. This is the celebrated Borel embedding, and was uniformly studied for symmetric R-spaces and generalized Grassmannians in [5]. The non-compact dual to $\mathrm{Gr}(n,d)$ is the "space-like Grassmannian" which can be thought of as a generalization of hyperbolic space.

We generalize this to framed quiver varieties. The key idea is that different types of quiver varieties will arise by considering space-like representations with respect to different choices of quadratic forms on the framing. As explained in the Introduction, our motivation is to find a relation between our formulation of neural networks and the original Euclidean formulation. Using this construction, we not only get an interpolation between these two different formulations, but also find non-compact type quiver varieties which can also be used in machine learning. Such a family of quiver varieties of different types is what we are referring to as the uniformization of framed quiver varieties mentioned in the title.

4.1. A quick review

Let Q be a directed graph. Denote by Q_0, Q_1 the set of vertices and arrows respectively. A quiver representation w with dimension vector $d \in \mathbb{Z}_{\geq 0}^{Q_0}$ associates each arrow a with a matrix w_a of size $d_{h(a)} \times d_{t(a)}$ (where $h(a), t(a)$ denote the head and tail vertices of a respectively). The set of complex quiver representations with dimension $\vec{d}$ form a vector space denoted by $R_{\vec{d}}(Q)$. $\mathrm{GL}(d) := \prod_{i \in Q_0} \mathrm{GL}(d_i, \mathbb{C})$ acts on $R_{\vec{d}}(Q)$ via

$$g \cdot (w_a : a \in Q_1) = (g_{h(a)} \cdot w_a \cdot g_{t(a)}^{-1} : a \in Q_1). \tag{18}$$

Let $d, n \in \mathbb{Z}_{\geq 0}^{Q_0}$. n will be the dimension vector for the framing, which is a linear map $e^{(i)} : \mathbb{C}^{n_i} \to V_i$ at each $i \in Q_0$ (where $V_i = \mathbb{C}^{d_i}$).

Theorem 4.1 ([23]). *The vector space of framed representations is given by*

$$R_{n,d} = R_d \times \bigoplus_{i \in Q_0} \mathrm{Hom}(\mathbb{C}^{n_i}, \mathbb{C}^{d_i}).$$

It carries a natural action of $\mathrm{GL}(d)$ *given by* $g \cdot (w, e) = (g \cdot w, (ge^{(i)} : i \in Q_0))$, *where* $g \cdot V$ *is given by Equation* (18). $(w, e) \in R_{n,d}$ *is called stable if there is no proper subrepresentation* U *of* w *which contains* $\mathrm{Im}\, e$. *The set of all stable points of* $R_{n,d}$ *is denoted by* $R_{n,d}^s$. *Then the quotient* $\mathcal{M}_{n,d} := R_{n,d}^s/\mathrm{GL}(d)$ *is a smooth variety, which is called a framed quiver moduli.*

The topology of $\mathcal{M}_{n,d}$ is well-understood. Let's make an ordering of the vertices. Namely the vertices are labeled by $\{1, \ldots, N\}$, such that $i < j$ implies there is no arrow going from j to i. Such a labeling exists if Q has no oriented cycle.

Theorem 4.2 (Reineke [26]). *Assume* Q *has no oriented cycle. Consider the chain of iterated Grassmannian bundles* $M^{(N)} \xrightarrow{p_N} M^{(N-1)} \xrightarrow{p_{N-1}} \ldots \xrightarrow{p_2} M^{(1)} \xrightarrow{p_1} \mathrm{pt}$ *(where* pt *denotes a singleton) defined by induction:*

$$M^{(i)} = Gr_{M^{(i-1)}} \left(\underline{\mathbb{C}^{n_i}} \oplus \bigoplus_{j \to i} p_{i-1}^* \ldots p_{j+1}^*(S_j), d_i \right) \to M^{(i-1)},$$

where S_i *denotes the tautological bundle on* M_i *(as a Grassmannian bundle over* M_{i-1}*). (The direct sum is over each arrow* $j \to i$.*) Then* $\mathcal{M}_{\vec{n},\vec{d}} \cong M^{(N)}$, *with universal bundles* $\mathcal{V}_i \cong p_N^* \ldots p_{i+1}^* S_i$ *for all* $i \in Q_0$.

In the previous paper [19] we introduced a Hermitian metric H_i for each of these $\mathcal{V}_i$ and showed that its Ricci curvature induces a Kähler metric on $\mathcal{M}$. Let's quickly review this construction.

Theorem 4.3 ([19]). *Let* Q *be a finite quiver. Let* $R_{n,d}$ *be the space of framed quiver representations of* Q *with representing dimension* d *and framing dimension* n. *For any path* γ *in* Q, *let* $e^{t(\gamma)}$ *be the framing map associated to the vertex* $t(\gamma)$ *and let* w_γ *be the matrix representation of* γ.

For a fixed vertex (i), *let* ρ_i *be the row vector whose entries are all the elements of the form* $w_\gamma e^{t(\gamma)} : R_{n,d} \to \mathrm{Hom}(\mathbb{C}^{n_{t(\gamma)}}, \mathbb{C}^{d_i})$ *such that* $h(\gamma) = i$. *Consider*

$$\rho_i \rho_i^* = \sum_{h(\gamma)=i} \left(w_\gamma e^{t(\gamma)} \right) \left(w_\gamma e^{t(\gamma)} \right)^* \tag{19}$$

as a map $\rho_i\rho_i^* : R_{n,d} \to End(\mathbb{C}^{d_i})$.

Then $(\rho_i\rho_i^*)^{-1}$ *is* $\mathrm{GL}(d)$*-equivariant and descends to a Hermitian metric on* $\mathcal{V}_i$ *over* $\mathcal{M}$*. We denote this resulting metric as* H_i*.*

Suppose Q *has no oriented cycle. Then*

(20)
$$H_T := \sum_i \partial\bar{\partial} \log\det H_i = \sum_i \left(tr(\partial\rho_i)^* H_i \partial\rho_i - tr\left(H_i\rho_i(\partial\rho_i)^* H_i(\partial\rho_i)\rho_i^*\right)\right)$$

defines a Kähler metric on $\mathcal{M}$*.*

Consider the simplest possible example, a quiver with a single vertex.

Example 4.4. *Let* Q *consist of a single vertex* (1) *with no arrows. Let the representing dimension and the framing dimension be* d *and* n *respectively where* $d < n$*. The framed quiver moduli is simply* $Gr(n,d)$*, the (dual) Grassmannian of surjective linear maps* $\mathbb{C}^n \to \mathbb{C}^d$*. Equation 19 becomes* $H = ee^*$ *on the universal bundle over the dual Grassmannian. If we take the chart where the first* d*-many components of* e *form an invertible map, we can rewrite* e *as* $e = (Id_d, b)$ *due to the* G_d*-equivalence. Then* H *becomes* $(Id_d + b)^{-1}$*, the standard metric on the universal bundle over* $Gr(n,d)$*. In particular, for* $d = 1$*,* $Gr(n,1)$ *is the projective space* $\mathbb{P}^{n-1}$*, and the Ricci curvature of* H *is the Fubini-Study metric.*

4.2. The non-compact dual of framed quiver moduli

Assume that $n_i \geq d_i \ \forall i$. We write the framing map as $e^{(i)} = (\epsilon_i \ b_i)$ where ϵ_i and b_i are respectively the "basis part" and "bias part of our framing map $e^{(i)}$. Then Equation 19 can be modified to be:

$$H_i^\alpha = \left(\epsilon_i\epsilon_i^* + \alpha b_i b_i^* + \sum_{\gamma: h(\gamma)=i, \gamma\neq 1} \alpha_\gamma w_\gamma e^{t(\gamma)} \left(w_\gamma e^{t(\gamma)}\right)^*\right)^{-1}. \tag{21}$$

Here, $\gamma \neq 1$ means γ not equal to the trivial path, as the first two terms already account for that option.

It is this generalization of the metric which we use for the uniformization. By varying α and α_γ, we get different quadratic forms. For example, in Equation 19, α and all α_γ are simply 1. The zero curvature case will elaborated on later in Section 4.3.

Remark 4.5. *The application of hyperbolic geometry has mostly focused on fiber direction in existing literature, namely the representation spaces (and*

their corresponding universal bundles over the moduli). Here, we are concerned about metrics on the moduli space (playing the role of the weight space). This is general for all quiver moduli, not just restricted to specific models. Thus, in this moduli approach, the method of varying metrics (with positive, zero or negative curvatures) can be applied to any model of machine learning.

For now we will set the α and α_γ to -1 to consider the negative curvature case. Namely,

$$H_i^- := \left(\epsilon_i \epsilon_i^* - b_i b_i^* - \sum_{\gamma : h(\gamma)=i, \gamma \neq 1} w_\gamma e^{t(\gamma)} \left(w_\gamma e^{t(\gamma)} \right)^* \right)^{-1}. \tag{22}$$

It must be emphasized that this quadratic form is **not** positive-definite on $\mathcal{V}_i$ and thus cannot serve as a metric.

The main idea here is that we restrict to the subset of the moduli space where this quadratic form is positive-definite and thus gives a metric. This restriction gives the non-compact dual of the framed quiver moduli. As before, let's consider the A_1-quiver and what this metric looks like on that quiver in particular.

Example 4.6. *Let Q consist of a single vertex with no arrows. Let the framing dimension of the representation space be n, and suppose the representing dimension at the single vertex is 1. Equation 22 becomes*

$$H^- = (|\epsilon|^2 - |b|^2)^{-1}.$$

Since we restrict to the subset where H^- is positive-definite, ϵ needs to be nonzero. By applying the quiver automorphism, ϵ can be rescaled to be 1. Thus $H^- = (1 - |b|^2)^{-1}$, and $|b|^2 < 1$. This gives the hyperbolic moduli, which is the open unit ball in $\mathbb{C}^{n-1}$. The Ricci curvature of H^- gives the Poincaré metric.

Thus from Examples 4.4 and 4.6 we can see the motivating duality mentioned at the start of the section.

Definition 4.7. *Assume Q has no oriented cycle. Let ρ_i be as in Theorem 4.3 so that ρ_i is a row vector with entries of the form $w_\gamma e^{t(\gamma)}$ where γ is some path in Q ending at vertex (i), w_γ is the representing matrix of this path, and $e^{t(\gamma)}$ is the framing map at $t(\gamma)$, the starting vertex of γ. Arrange the entries*

of ρ_i so that the first n_i-many entries correspond to the framing arrows at vertex (i). Then let H_i^- be the quadratic form defined by:

$$H_i^- = \left(\rho_i \begin{pmatrix} I_{d_i} & 0 \\ 0 & -I_{N_i - d_i} \end{pmatrix} \rho_i^*\right)^{-1} \tag{23}$$

Here, $N_i = \sum_{\gamma : h(\gamma) = i} n_{t(\gamma)}$. We define $R_{n,d}^-$ to be the subset of $R_{n,d}$ where H_i^- is positive-definite for all i.

In particular, n_j gets counted once for each distinct path from (j) to (i). Note that $N_i \geq d_i$ for all i when $\mathcal{M} \neq \emptyset$, which we always assume to be the case. Unlike the intuition given at the start of the subsection, this definition does not require that $n_i \geq d_i$ for all i.

Proposition 4.8.

$$R_{n,d}^- \subset R_{n,d}^s.$$

Proof. Consider a point in $R_{n,d}^-$. Write $\rho_i = (\epsilon_i \ R)$ evaluated at this point as a $(d_i \times N_i)$-matrix, where ϵ_i is a $(d_i \times d_i)$-matrix and R is the remaining part. Then $H_i^- = (\epsilon_i \epsilon_i^* - RR^*)^{-1}$. We claim that ϵ_i must be invertible, and hence ρ_i is surjective. This is true for all i, and hence the point is stable.

Suppose ϵ_i is not invertible. Then there exists v such that $\epsilon_i^* \cdot v = 0$. Then $v^* H_i^- v = -v^* R R^* v \leq 0$, contradicting that H_i^- evaluated at each point in $R_{n,d}^-$ is positive-definite. □

Lemma 4.9. *H_i^- is G_d-equivariant and $R_{n,d}^-$ is G_d-invariant.*

Proof. H_i^- is $GL(d)$-equivariant because

$$\left((g \cdot w_\gamma e^{t(\gamma)})(g \cdot w_\gamma e^{t(\gamma)})^*\right)^{-1} = g^{-1} \left((w_\gamma e^{t(\gamma)})(w_\gamma e^{t(\gamma)})^*\right)^{-1} (g^*)^{-1}.$$

The reason that $R_{n,d}^-$ is $GL(d)$-invariant is because if

$$x^* \left((w_\gamma e^{t(\gamma)})(w_\gamma e^{t(\gamma)})^*\right)^{-1} x > 0,$$

then

$$(g \cdot x)^* \left((g \cdot w_\gamma e^{t(\gamma)})(g \cdot w_\gamma e^{t(\gamma)})^*\right)^{-1} (g \cdot x) = x^* \left((w_\gamma e^{t(\gamma)})(w_\gamma e^{t(\gamma)})^*\right)^{-1} x > 0$$

by the $GL(d)$-equivariance of H_i^-. Thus, action by $GL(d)$ sends $R_{n,d}^-$ to itself. □

Definition 4.10. *If $n_i \geq d_i$, then $e^{(i)}$ can be written as $e^{(i)} = (\epsilon_i, b_i)$ where ϵ_i is the d_i-many components of $e^{(i)}$ and b_i is the remaining $(n_i - d_i)$-many components. We call ϵ_i to be the basis part of $e^{(i)}$ and b_i to be the bias part of $e^{(i)}$.*

We call ϵ_i the basis part because we think of it as imposing a basis on V_i. Similarly, b_i is the bias part because in the standard case where $n_i = d_i + 1$, it encodes the translation bias parameter in the neural network sense.

From now on, we will assume $n_i \geq d_i$, which is the case in applications. This assumption also ensures that the choices of negative signs in defining H_i^- for different vertices i are compatible, so that $R^-_{n,d} \neq \emptyset$.

Proposition 4.11. *Assume that $n_i \geq d_i$ for all i.*

$$\emptyset \neq R^-_{n,d} \subset \{\epsilon_i \text{ is invertible for all } i\} \subset R^s_{n,d}.$$

Proof. From the proof of Proposition 4.8, it is clear that ϵ_i is invertible over $R^-_{n,d}$ and these points belong to $R^s_{n,d}$. To see that $R^-_{n,d} \neq \emptyset$, we can take $\epsilon_i = \mathrm{Id}$ and $b_i = 0$ for all $i \in Q_0$, and all the representing matrices for the arrows of Q to be 0. This gives a point in $R_{n,d}$ at which $H_i^- = \mathrm{Id}$ is positive-definite. □

Suppose a Lie group G acts on a vector bundle $V \xrightarrow{\pi} M$ equivariantly fiberwise linearly, and the action of G on M is free and proper. A metric H on V is G-equivariant if

$$H_x(v, w) = H_{g \cdot x}(g \cdot v, g \cdot w).$$

It is possible that V may not descend to a vector bundle over M/G if $G_p \subset G$ acts on V non-trivially at a point $p \in M$. In the case that the corresponding bundle does exist, H will descend to that bundle if and only if H is G-equivariant.

Since we know that $R^-_{n,d}$ is a $\mathrm{GL}(d)$-invariant non-compact open subset, we can quotient by $\mathrm{GL}(d)$ in the same way we do for $R^s_{n,d}$.

Definition 4.12. *We define the dual of $\mathcal{M}$ as the quotient $\mathcal{M}^- = R^-_{n,d}/\mathrm{GL}(d)$ with universal bundles $\mathcal{V}_i^- := (R^-_{n,d} \times \mathbb{C}^{d_i})/\mathrm{GL}(d_i)$. Since H_i^- is Hermitian and G_d-equivariant, it descends to a metric on $\mathcal{V}_i^-$ over $\mathcal{M}^-$.*

Remark 4.13. *As a result of Proposition 4.11 and the fact that $\mathrm{GL}(d)$ acts only on the left on the framing space, $e^{(i)} = (\epsilon_i, b_i) = (I_{d_i}, \tilde{b}_i)$ where $\tilde{b}_i = \epsilon_i^{-1} b_i$ and is itself a generic bias vector for each i. Thus, from this point forward we will be assuming both that $n_i \geq d_i$ for all i and that all framing maps are of the form $e^{(i)} = (\mathrm{Id}, b_i)$. Thus $\mathcal{M}^- \subset R_{n-d,d}$.*

Example 4.14. *Consider the framed A_1 quiver (the quiver with one vertex and zero arrows). Let $d \leq n = N$. Then $\mathcal{M}$ is $Gr(n,d)$. As a Hermitian symmetric space, this is dual to the space-like Grassmannian $Gr^-(n,d)$. Here, we define $Gr^-(n,d)$ to be the open subset of $Gr(n,d)$ consisting of d-planes in $\mathbb{C}^n$ where the quadratic form*

$$Q(x,y) = \sum_{i=1}^{d} \overline{x}_i y_i - \sum_{j=d+1}^{n} \overline{x}_j y_j$$

is positive-definite.

Similar to Remark 4.13, we can take elements of $Gr^-(n,d)$ to be of the form (Id, b) where the first d-many columns are the $d \times d$ identity matrix and b is the remaining $d \times (n-d)$ columns. Then we can say that $Gr^-(n,d)$ is the set $\{b \in \mathbb{C}^{d\times(n-d)} : \mathrm{Id} - bb^* \geq 0\}$.

Going back to the quiver, since there are no other arrows, we see that $H_1^- = (\mathrm{Id} - bb^*)^{-1}$. Thus, $\mathcal{M}^-$ is going to be the set $\{b : \mathrm{Id} - bb^* \geq 0\}$.

In particular, when $d = 1$, $Gr(n,1)^-$ is complex hyperbolic space and the Ricci curvature of $H_1^- = \frac{1}{1-|b|^2}$ is the standard metric for the Poincare disk model of complex hyperbolic space, just like in Example 4.6.

Now we define an explicit metric on $\mathcal{M}^-$, using (23) written in terms of paths in Q, in an analogous way as the one given in Theorem 4.3.

Theorem 4.15. *Assume Q is acyclic. Define $H_T^- := -i\sum_i \partial\overline{\partial} \log\det H_i^-$ on $\mathcal{M}^-$. Then H_T^- is a Kähler metric on $\mathcal{M}^-$.*

Proof. This proof is similar to that of Theorem 3.15 in [19]. We include the details for the reader's convenience.

Let's denote $\rho = \rho^{(i)} = \left(w_\gamma e^{(t(\gamma)}\right)_{\gamma:h(\gamma)=i}$ which is a matrix-valued function on $R_{n,d}^-$. At each point of $R_{n,d}^-$, we have that ρ is a linear map from $\hat{W}_i := \bigoplus_{\gamma:h(\gamma)=i} \mathbb{C}^{n_{t(\gamma)}}$ to V_i. The Ricci curvature of the metric H_i^- is given by $i\partial\overline{\partial}\log\det\rho A\rho^*$ where A is the matrix $\mathrm{diag}(1,-1,-1,\ldots,-1)$. Let B be the matrix $\mathrm{diag}(1,\sqrt{-1},\ldots,\sqrt{-1})$ and define $\hat{\rho} := \rho B$ so that $\hat{\rho}\hat{\rho}^* = \rho A\rho^*$. Thus we have that $H_i^- = (\hat{\rho}\hat{\rho}^*)^{-1}$.

We can take the singular valued decomposition of $\hat{\rho}$ to write it as

$$\hat{\rho} = U \cdot (\mathrm{diag}(\lambda_1,\ldots,\lambda_{d_i}) \quad 0) \cdot V^*$$

where $U \in U(d_i)$, $V \in U(\dim \hat{W}_i)$, and the λ_i are all positive real numbers. We know that none of the λ_i are zero since that would make corresponding

quiver representations non-surjective and thus unstable. Then

$$\hat{\rho} = U \cdot (\mathrm{diag}(\lambda_1, \lambda_2, \ldots, \lambda_{d_i}) \quad 0) \cdot V^*,$$

$$\hat{\rho}\hat{\rho}^* = U \left(\mathrm{diag}(\lambda_1^2, \lambda_2^2, \ldots, \lambda_{d_i}^2)\right) U^*,$$

$$\hat{\rho}^*(\hat{\rho}\hat{\rho}^*)^{-\frac{1}{2}} = V \begin{pmatrix} \mathrm{diag}(\lambda_1, \ldots, \lambda_{d_i}) \\ 0 \end{pmatrix} (\mathrm{diag}(\lambda_1^{-1}, \ldots, \lambda_{d_i}^{-1}) U^* = V \begin{pmatrix} I_{d_i} \\ 0 \end{pmatrix} U^*.$$

Let us consider the decomposition $\hat{W}_i = (\mathrm{Im}\,\hat{\rho}^*) \oplus (\mathrm{Im}\,\hat{\rho}^*)^{\perp}$. In particular, this shows that $\hat{\rho}^*(\hat{\rho}\hat{\rho}^*)^{-\frac{1}{2}}$ is the orthogonal embedding of V_i to $\mathrm{Im}\,\hat{\rho}^* \subset \hat{W}_i$.

Then

$$\partial\overline{\partial} \log\det \hat{\rho}\hat{\rho}^* = \partial \left(\mathrm{tr}\, \left((\hat{\rho}\hat{\rho}^*)^{-1}\overline{\partial}(\hat{\rho}\hat{\rho}^*)\right)\right)$$

$$= \mathrm{tr}\, \left(\partial \left((\hat{\rho}\hat{\rho}^*)^{-1}(\hat{\rho})(\partial\hat{\rho})^*\right)\right)$$

$$= \mathrm{tr}\, \left((\hat{\rho}\hat{\rho}^*)^{-1}(\partial\hat{\rho})(\partial\hat{\rho})^* + \left(\partial(\hat{\rho}\hat{\rho}^*)^{-1}\right)\hat{\rho}(\partial\hat{\rho})^*\right)$$

$$= \mathrm{tr}\, \left((\partial\hat{\rho})^*(\hat{\rho}\hat{\rho}^*)^{-1}(\partial\hat{\rho})\right) - \mathrm{tr}\, \left((\hat{\rho}\hat{\rho}^*)^{-1}(\partial(\hat{\rho}\hat{\rho}^*))(\hat{\rho}\hat{\rho}^*)^{-1}\hat{\rho}(\partial\hat{\rho})^*\right)$$

$$= \mathrm{tr}\, \left((\partial\hat{\rho})^*(\hat{\rho}\hat{\rho}^*)^{-1}(\partial\hat{\rho})\right) - \mathrm{tr}\, \left((\hat{\rho}\hat{\rho}^*)^{-1}\hat{\rho}(\partial\hat{\rho})^*(\hat{\rho}\hat{\rho}^*)^{-1}(\partial\hat{\rho})\hat{\rho}^*\right)$$

$$= \mathrm{tr}\, \left((\partial\hat{\rho})^*(\hat{\rho}\hat{\rho}^*)^{-1}(\partial\hat{\rho})\right)$$

$$-\mathrm{tr}\, \left(\left((\partial\hat{\rho}) \cdot \left(\hat{\rho}^*(\hat{\rho}\hat{\rho}^*)^{-\frac{1}{2}}\right)\right)^* (\rho\rho^*)^{-1} \left((\partial\hat{\rho}) \cdot \left(\hat{\rho}^*(\hat{\rho}\hat{\rho}^*)^{-\frac{1}{2}}\right)\right)\right).$$

Consider a vector $v \in T^{1,0}R^-_{n,d} \cong TR^-_{n,d}$. We can see that the term $\mathrm{tr}\, \left((\partial_v\hat{\rho})^*(\hat{\rho}\hat{\rho}^*)^{-1}(\partial_v\hat{\rho})\right)$ is in fact the square norm of the linear map $\partial_v\hat{\rho}$ with respect to the metric H_i^-. Using the decomposition of $\hat{W}_i$ above, let's write $\partial_v\hat{\rho}$ as the decomposition $\partial_v\hat{\rho} = ((\partial_v\hat{\rho})_1, (\partial_v\hat{\rho})_2)$ where $(\partial_v\hat{\rho})_1 : \mathrm{Im}\,\hat{\rho}^* \to V_i$ and $(\partial_v\hat{\rho})_2 : (\mathrm{Im}\,\hat{\rho}^*)^{\perp} \to V_i$. In particular, given the previous discussion, we see that $(\partial_v\hat{\rho})_1$ is actually $\partial_v\hat{\rho}$ composed with $\hat{\rho}^*(\hat{\rho}\hat{\rho}^*)^{-\frac{1}{2}}$. Thus, we can see that the other term

$$\mathrm{tr}\, \left(\left((\partial_v\hat{\rho}) \cdot \left((\hat{\rho})^*(\hat{\rho}\hat{\rho}^*)^{-\frac{1}{2}}\right)\right)^* (\hat{\rho}\hat{\rho}^*)^{-1} \left((\partial_v\hat{\rho}) \cdot \left((\hat{\rho})^*(\hat{\rho}\hat{\rho}^*)^{-\frac{1}{2}}\right)\right)\right)$$

is actually the square norm of $\partial\hat{\rho}_1$ (with respect to the H_i^- metric). Then we have

$$i\partial\overline{\partial} \log\det H_i^- = |\partial\hat{\rho}|_{H_i^-} - |\partial\hat{\rho}_1|_{H_i^-} = |\partial\hat{\rho}_2|_{H_i^-}.$$

Thus, the Ricci curvature is semi-positive definite.

Now suppose $\left(\partial_v \hat{\rho}^{(i)}\right)_2 = 0$ for all i. Then the image of $\left(\partial_v \hat{\rho}^{(i)}\right)^* = \partial_v(\hat{\rho}^{(i)})^*$ is in the image of $(\hat{\rho}^{(i)})^*$. Thus ∂_v does not alter the subspaces given by $(\hat{\rho}^{(i)})^* : V_i \to \hat{W}_i$. $((\hat{\rho}^{(i)})^*)_{i\in I}$ gives an embedding of $\mathcal{M}^-$ to the product of Grassmannians of subspaces in $\hat{W}_i$. Since ∂_v does not change the subspaces, it must be the zero tangent vector. As a result, the curvature is positive definite and defines a Kähler metric. □

Example 4.16. *Consider the framed A_2 quiver. This quiver has vertices (1) and (2) and has one arrow a going from (1) to (2). Then*

$$H_1^- = (Id_{d_1} - b_1 b_1^*)^{-1}$$

and

$$H_2^- = (Id_{d_2} - b_2 b_2^* - w_a w_a^* - w_a b_1 b_1^* w_a^*)^{-1} = (Id_{d_2} - b_2 b_2^* - w_a H_1^{-1} w_a^*)^{-1}$$

where $H_1 = (\mathrm{Id}_{d_1} + b_1 b_1^*)^{-1}$ is the Hermitian metric on $\mathcal{V}_1$ in Definition 4.3.

Using Gram-Schmidt orthonormalization, we can write $H_1^{-1} = g(b_1)g(b_1)^*$ for some $g(b_1) \in \mathrm{GL}(d_1)$. Thus

$$H_2^- = (Id_{d_2} - (w_a g(b_1))(w_a g(b_1))^* - b_2 b_2^*)^{-1}.$$

$\mathcal{M}^- = \{(b_1, w_a, b_2) : H_1^-$ and H_2^- are positive definite$\}$. Then we have the map

$$\mathcal{M}^- \to Gr(n_1, d_1)^- \times Gr(n_2 + d_1, d_2)^-$$

by $(b_1, w_a, b_2) \mapsto (b_1, (w_a g(b_1), b_2))$, which is invertible. We have identifications of the universal bundles $(\mathcal{V}_i, H_i^-)$ with the pullback of tautological bundles over $Gr(n_1, d_1)^-$ and $Gr(n_2 + d_1, d_2)^-$ respectively, which are compatible with this diffeomorphism.

We can go much further than this. In fact, for general acyclic quivers there exists an identification between $(\mathcal{V}_i^-, H_i^-)$ over $\mathcal{M}^-$ and the tautological bundles over space-like Grassmannians as in the above example, if we ignore complex structures.

Theorem 4.17. *Assume that the underlying quiver Q is acyclic. Then there exists a symplectomorphism*

$$\phi : (\mathcal{M}^-, H^-_{T\mathcal{M}^-}) \overset{\cong}{\to} \prod_i (Gr^-(m_i, d_i), H^-_{(m_i,d_i)})$$

that restricts to a diffeomorphism between the real loci, and a bundle isomorphism

$$(\mathcal{V}_i^-, H_i^-) \xrightarrow{\cong} (\phi^* U_i, H^-_{(m_i,d_i)})$$

that restricts to a bundle isomorphism between the corresponding real vector bundles over the real loci. Here U_i is the tautological bundle over $Gr^-(m_i, d_i)$, $m_i = n_i + \sum_{a:h(a)=i} d_{t(a)}$, and $H^-_{(m_i,d_i)}$ is the standard metric of $Gr^-(m_i, d_i)$.

First, we make the following lemma.

Lemma 4.18. $H_i^- = \left(\mathrm{Id} - b_i b_i^* - \sum_{a:h(a)=i} w_a H_{t(a)}^{-1} w_a^*\right)^{-1}$.

Proof. Consider vertex (i) in quiver Q. Let's denote $\Gamma_i := \{\gamma : h(\gamma) = i\}$, the paths ending at (i).

Aside from the trivial path, every γ in Γ_i must be of the form $a \cdot \gamma$ for some arrow a with $h(a) = i$. Thus, we can decompose $\Gamma_i = \{(i)\} \cup \bigcup_{a:h(a)=i} a \cdot \Gamma_{t(a)}$ where (i) denotes the trivial path. Because of this, we can write

$$H_i^- = \left(\mathrm{Id} - b_i b_i^* - \sum_{a:h(a)=i} \sum_{\gamma \in \Gamma_{t(a)}} w_{a\cdot\gamma} e^{t(a\cdot\gamma)} \left(w_{a\cdot\gamma} e^{t(a\cdot\gamma)}\right)^*\right)^{-1}.$$

We have $w_{a\cdot\gamma} = w_a \cdot w_\gamma$ where w_a is the linear map associated to the arrow a. Moreover, $t(a\gamma) = t(\gamma)$. Thus

$$\begin{aligned} H_i^- &= \left(\mathrm{Id} - b_i b_i^* - \sum_{a:h(a)=i} \sum_{\gamma \in \Gamma_{t(a)}} w_a w_\gamma e^{t(\gamma)} \left(w_a w_\gamma e^{t(\gamma)}\right)^*\right)^{-1} \\ &= \left(\mathrm{Id} - b_i b_i^* - \sum_{a:h(a)=i} w_a \left(\sum_{\gamma \in \Gamma_{t(a)}} w_\gamma e^{t(\gamma)} \left(w_\gamma e^{t(\gamma)}\right)^*\right) w_a^*\right)^{-1} \\ &= \left(\mathrm{Id} - b_i b_i^* - \sum_{a:h(a)=i} w_a H_j^{-1} w_a^*\right)^{-1}. \end{aligned}$$

□

Proof of Theorem 4.17. Let (i) be a vertex. By Lemma 4.18, we can write H_i^- as

$$\mathrm{Id} - b_i b_i^* - \sum_{a:h(a)=i} w_a H_{t(a)}^{-1} w_a^*.$$

By Gram-Schmidt normalization, we can write $H_{t(a)}^{-1} = g_{t(a)} g_{t(a)}^*$ for some $g_{t(a)} \in \mathrm{GL}(d_{t(a)})$. Then

$$H_i^- = \left(\mathrm{Id} - b_i b_i^* - \sum_{a:h(a)=i} w_a g_{t(a)} g_{t(a)}^* w_a^* \right)^{-1} = (\mathrm{Id} - ww^*)^{-1}$$

where

$$w = b_i \oplus \bigoplus_{a:h(a)=i} w_a g_{t(a)}.$$

Thus, we define $\phi : (\mathcal{M}^-, H^-_{T\mathcal{M}^-}) \to \prod_i (\mathrm{Gr}^-(m_i, d_i), H^-_{(m_i,d_i)})$ by sending $(b_i, w_a)_{i \in Q_0, a \in Q_1}$ to $(b_i, (w_a g_{t(a)})_{a:h(a)=i})_{i \in Q_0}$.

The map ϕ is invertible: for each $i \in Q_0$, g_i only depends on b_j for $j \in Q_0^{(i)}$ and w_a for $a \in Q_1^{(i)}$, where $Q^{(i)}$ is the sub-quiver containing those arrows that can be a part of a path heading to i. We can invert ϕ inductively as follows. For fixed vertex i, if $Q_0^{(i)}$ is empty, then $Q_1^{(i)}$ is also empty, so ϕ acts trivially by sending b_i to itself. If $Q_0^{(i)}$ is not empty, then for fixed arrow $a \in Q_1^{(i)}$, ϕ sends (b_i, w_a) to (b_i, w_ϕ) where $w_\phi = w_a g_{t(a)}$. By construction,

$$(w_\phi)(w_\phi)^* = (w_a g_{t(a)})(w_a g_{t(a)})^* = w_a H^-_{t(a)} w_a^*.$$

If we know $H^-_{t(a)}$, then we can calculate $g_{t(a)}$ and thus recover w_a. This reduces the problem to a calculation on $Q^{t(a)}$. Since Q is acyclic, repeating this process inductively is guaranteed to eventually reduce down to source vertices (meaning vertices k such that $Q_0^{(k)}$ is empty). In particular, $H_k^- = (\mathrm{Id} - b_k b_k^*)^{-1}$ for all source vertices k.

Since ϕ identifies H_i^- with the standard metric on the tautological bundle of $\mathrm{Gr}^-(m_i, d_i)$, and the symplectic form is $H_T^- = -i \sum_i \partial\overline{\partial} \log\det H_i^-$, ϕ is a symplectomorphism. Written in these coordinates, the map $(\mathcal{V}_i^-, H_i^-) \xrightarrow{\cong} (\phi^* U_i, H_{(m_i,d_i)})$ is simply given by identity.

Restricting to b_i and w_a having real coordinates, g_i produced from the Gram-Schmidt process is a real matrix. Thus ϕ restricts as a diffeomorphism between the real loci. □

Remark 4.19. *This correspondence between $\mathcal{M}^-$ and $Gr^-(m_i, d_i), H^-_{(m_i,d_i)}$ is **only** a symplectomorphism, since the Gram-Schmidt process is not holomorphic.*

4.3. Euclidean signature

In addition to the non-compact dual, we can use Equation 21 to get other moduli spaces in the same vein. The most straightforward variant is achieved by setting α and all of the α_γ to zero. This means throwing out the contribution coming from anything other than the first d_i-many framing arrows.

Definition 4.20. *Assume Q has no oriented cycle. Let ρ_i be as in definition 4.3 so that ρ_i is a row vector with entries of the form $w_\gamma e^{t(\gamma)}$ where γ is some path in Q ending at vertex (i), w_γ is the representing matrix of this path, and $e^{t(\gamma)}$ is the framing map at $t(\gamma)$, the starting vertex of γ. Arrange the entries of ρ_i so that the first n_i-many entries correspond to the framing arrows at vertex (i). Then let H_i^0 be the quadratic form defined by:*

$$H_i^0 = \left(\rho_i \begin{pmatrix} I_{d_i} & 0 \\ 0 & 0 \end{pmatrix} \rho_i^*\right)^{-1} \tag{24}$$

Here, $N_i = \sum_{\gamma: h(\gamma)=i} n_{t(\gamma)}$. We define $R_{n,d}^0$ to be the subset of $R_{n,d}$ where H_i^0 is positive-definite for all i.

Note that we still need $N_i \geq d_i \forall i$ to have $\mathcal{M} \neq \emptyset$, thus we will still be assuming that to be the case. Indeed, most of the following statements are copied or follow from analogous statements in Section 4.2.

Proposition 4.21. *$R_{n,d}^0 \subset R_{n,d}^s$.*

Proof. As in Proposition 4.8, consider a point in $R_{n,d}^0$. Write $\rho_i = (\epsilon_i \; R)$ evaluated at this point as a $(d_i \times N_i)$-matrix, where ϵ_i is a $(d_i \times d_i)$-matrix and R is the remaining part. Then $H_i^0 = (\epsilon_i \epsilon_i^*)^{-1}$. If ϵ_i is not invertible, then $\epsilon_i \epsilon_i^*$ is not positive-definite. Thus, for a point in $R_{n,d}^0$, we have that ϵ_i is invertible for all i which means that ρ_i is surjective for all i. Thus the point is stable. □

Lemma 4.22. *H_i^0 is G_d-equivariant and $R_{n,d}^0$ is G_d-invariant.*

Proof. This follows directly from Lemma 4.9.

□

Similar to Section 4.2, we will assume $n_i \geq d_i$ from this point forward. Thus, we can talk about the framing part ϵ_i of $e^{(i)}$ corresponding to the first d_i-many components, and the bias part b_i of $e^{(i)}$ corresponding to the remaining $(n_i - d_i)$-many components. With this, H_i^0 can be written simply as

$$H_i^0 = (\epsilon_i \epsilon_i^*)^{-1}$$

Proposition 4.23. *Assume that $n_i \geq d_i$ for all i.*

$$\emptyset \neq R^0_{n,d} = \{\epsilon_i \text{ is invertible for all } i\} \subset R^s_{n,d}.$$

Proof. From the proof of Proposition 4.21, it is clear that ϵ_i is invertible over $R^0_{n,d}$ and these points belong to $R^s_{n,d}$. Moreover, let w be any point of $R^s_{n,d}$ such that the framing parts ϵ_i of the framing maps $e^{(i)}$ are all invertible. Since H^0_i is only defined using ϵ_i, we can see that $w \in R^0_{n,d}$. Thus, $R^0_{n,d}$ is the subset of $R^s_{n,d}$ of points where the framing part is invertible. To see that $R^0_{n,d} \neq \emptyset$, we can take $\epsilon_i = \mathrm{Id}$ for all $i \in Q_0$ and set the remaining arrows to be zero. This gives a point in $R_{n,d}$ at which $H^0_i = \mathrm{Id}$ is positive-definite. □

Similar to $R^-_{n,d}$, since we know that $R^0_{n,d}$ is a $\mathrm{GL}(d)$-invariant non-compact open subset of $R^s_{n,d}$, we can directly quotient by $\mathrm{GL}(d)$.

Definition 4.24. *We define the Euclidean restriction of $\mathcal{M}$ as the quotient $\mathcal{M}^0 = R^0_{n,d}/\mathrm{GL}(d)$ with universal bundles $\mathcal{V}^0_i := (R^0_{n,d} \times \mathbb{C}^{d_i})/\mathrm{GL}(d_i)$. Since H^0_i is Hermitian and G_d-equivariant, it descends to a metric on $\mathcal{V}^0_i$ over $\mathcal{M}^0$.*

As a result of Proposition 4.23 and the fact that $\mathrm{GL}(d)$ acts only on the left on the framing space, $e^{(i)} = (\epsilon_i, b_i) = (I_{d_i}, \tilde{b}_i)$ where $\tilde{b}_i = \epsilon_i^{-1} b_i$ and is itself a generic bias vector for each i. Thus, from this point forward we will be assuming both that $n_i \geq d_i$ for all i and that all framing maps are of the form $e^{(i)} = (\mathrm{Id}, b_i)$. Thus $\mathcal{M}^0 \cong R_{n-d,d}$ and H^0_i can be taken to be the trivial metric on $\mathbb{C}^{d_i}$ for each i.

Over $\mathcal{M}^0$, activation functions have the simplest possible definition:

smooth (or piece-wise smooth) maps from $\mathbb{C}^{d_i}$ to itself. Any of the standard activation functions used in machine learning (sigmoid, ReLu, softmax, etc.) fit directly into this Euclidean restriction setting without any further modification.

Corollary 4.25. *H^0_i is the trivial metric on $\mathbb{C}^{d_i}$. Thus, $H^0_T := \sum_i \partial\overline{\partial} \log\det H^0_i$ is a Ricci-flat Kähler-Einstein metric and $R^-_{n,d} \subset R^0_{n,d}$.*

Remark 4.26. *Consider an acyclic quiver Q with dimension vector (n, d) such that $n_i = d_i$ for all source and sink vertices i, and $n_i = d_i + 1$ for all others. If Q with dimension vector (n, d) gives the underlying neuron structure for a neural network, then $\mathcal{M}^0$ is the training space for this network. In particular, the standard backward propagation algorithm for a feed-forward neural network is standard gradient descent in the relevant vector space, matching up exactly with the gradient descent on $\mathcal{M}^0$ induced by H^0_i.*

4.4. Hyperbolic activation functions

This point of view of uniformization provides a learning model over hyperbolic moduli, or more generally, interpolations of spherical, Euclidean and hyperbolic moduli. (One can add learnable parameters in the Hermitian metrics H_i, interpolating the metrics of different types.) This is hyperbolic learning in the base (that is, the parameter space). There is another direction that we can consider hyperbolic learning, namely the fiber bundle direction.

Recall that we have the universal vector bundles $\mathcal{V}_i$. In [19], we constructed activation function (as a fiber bundle map of $\mathcal{V}_i$) by composing the following:

$$\mathcal{V}_i \overset{H_i}{\cong} \mathcal{V}_i^* \overset{(e^{(i)})^*}{\to} \underline{\mathbb{C}^{n_i}} \overset{\sigma}{\to} \underline{\mathbb{C}^{n_i}} \overset{e^{(i)}}{\to} \mathcal{V}_i$$

where $\sigma : \mathbb{C}^{n_i} \to \mathbb{C}^{n_i}$ is a continuous function. We can do the same thing uniformly for $\mathcal{M}, \mathcal{M}^0$ and $\mathcal{M}^-$.

Additionally, in [19], we constructed a specific activation function as a symplectomorphism $(\mathbb{C}^n, \omega_{\mathbb{P}^n}|_{\mathbb{C}^n}) \cong (B, \omega_{\text{std}})$, where $B \subset \mathbb{C}^n$ is the ball $\{\|\vec{z}\|^2 < 1\}$, $\omega_{\mathbb{P}^n}$ is the Fubini-Study metric on $\mathbb{P}^n$, and ω_{std} is the standard symplectic form of $\mathbb{C}^n$. This symplectomorphism σ has the expression

$$(z_1, \dots, z_n) \to \left(\frac{z_1}{\sqrt{1 + \sum_{i=1}^n |z_i|^2}}, \dots, \frac{z_n}{\sqrt{1 + \sum_{i=1}^n |z_i|^2}} \right).$$

In view of hyperbolic metrics, we provide an alternative interpretation of the same function here.

Proposition 4.27. *σ gives a symplectomorphism $(\mathbb{C}^n, \omega_{std}) \to (\mathbb{CH}^n, \omega_{\mathbb{CH}^n})$ where $\mathbb{CH}^n$ denotes the hyperbolic ball.*

Proof. By definition, $\omega_{\mathbb{CH}^n}$ is equal to $-\partial\overline{\partial}\log(1-|w|^2)$ up to a simple scaling. Here, we will be thinking of w as the row vector $(w_1, \dots, w_n)$. Then

$$-\partial\overline{\partial}\log(1-|w|^2) = \partial\frac{wdw^*}{1-|w|^2}$$

$$= \frac{(1-|w|^2)dw \wedge dw^* + (dw \cdot w^*)wdw^*}{(1-|w|^2)^2} = \frac{(1-|w|^2)dw \wedge dw^* + \overline{w}dw^t\overline{dw}w^t}{(1-|w|^2)^2}$$

Now, let's similarly write z as the row vector $(z_1, \dots, z_n)$. We compute the pullback as

$$\sigma^*(dz \wedge dz^*) = d\frac{z}{\sqrt{1-|z|^2}} \wedge d\frac{z^*}{\sqrt{1-|z|^2}}$$

$$= \frac{(1-zz^*)dz + \frac{1}{2}(\overline{zdz^*} + zdz^*)z}{(1-zz^*)^{3/2}} \wedge \frac{(1-zz^*)dz^* + \frac{1}{2}(\overline{zdz^*} + zdz^*)z^*}{(1-zz^*)^{3/2}}$$

$$= \frac{1}{(1-zz^*)^3}\left((1-zz^*)^2 dz \wedge dz^* + \frac{1}{4}(\overline{zdz^*} + zdz^*)^2 zz^*\right)$$

$$+\frac{1}{(1-zz^*)^3}\left(\frac{1}{2}(1-zz^*)\left((\overline{zdz^*} + zdz^*)z \wedge dz^* + dz \wedge (\overline{zdz^*} + zdz^*)z^*\right)\right)$$

At this point, $\overline{zdz^*} \wedge zdz^* + dzz^* \wedge zdz^*$ can be rewritten as $2dz^t\overline{dz}z^t$. This gives us

$$= \frac{(1-zz^*)dz \wedge dz^* + \overline{z}dz^t\overline{d}zz^t}{(1-zz^*)^2}$$

which equals the above. □

In other words, σ gives an identification between $\mathbb{C}^n$ and $\mathbb{CH}^n$. Then signal propagation between hyperbolic spaces can be modeled simply as linear maps between $\mathbb{C}^n$, and the composition $\iota \circ \sigma : \mathbb{C}^n \to \mathbb{C}^n$, where $\iota : \mathbb{CH}^n \to \mathbb{C}^n$ is the inclusion of a ball in the space, gives an activation function.

4.5. Concrete implementation

In this subsection, we explain concrete formulas that can be directly coded in computer programs as well as some practical simplifications.

Let's first focus on a vertex i (a neuron). Equation (21) gives a family of metrics $H_i(\alpha_i)$ on the bundle $\mathcal{V}_i$. We take the chart in which ϵ_i is simply the identity matrix.

We can make the simplification that the representing dimensions over all vertices are one so that $H_i(\alpha_i)$ is just a (1×1)-matrix. There is a simple procedure called Abelianization which constructs a quiver with representing dimensions $d_i = 1$ from a quiver with a general dimension vector. Namely, if the representing dimension over vertex i is $d_i > 1$, we can split the vertex i into d_i-many vertices $\{(i,1), \ldots, (i,d_i)\}$ and set the representing dimension at each vertex (i,j) to be 1. For instance, Abelianization changes the A_3-quiver on the left of Figure 1 with general representing dimensions to the quiver in Figure 2.

To further simplify, the parameters α_γ in Equation 21 can be set to zero for long paths γ. For instance, we can set $\alpha_\gamma = 0$ for paths that consist of more than one arrows, and simplify Equation 21 to

$$H_i(\alpha_i) = \left(1 + \alpha_i(|\vec{w}_i|^2 + |b_i|^2)\right)^{-1} \tag{25}$$

where the α_i are (possibly learnable) parameters, $\vec{w}_i$ are the weights of the arrows heading to the vertex i, and b_i is the bias. In our algorithm, the activation at the i-th neuron is

$$\sigma_i\left(H_i(\alpha_i)\left(\vec{w}_i \cdot v_{\text{in}} + b_i\right)\right) = \sigma_i\left(\frac{\vec{w}_i \cdot v_{\text{in}} + b_i}{1 + \alpha_i(|\vec{w}_i|^2 + |b_i|^2)}\right). \tag{26}$$

We can understand H_i as filters depending on the parameters α_i that are concerned with the importance of the i-neuron. Setting $\alpha_i = 0$ reduces to the original algorithm without H_i. Moreover, the expressions $\vec{w}_i/(1 + \alpha_i(|\vec{w}_i|^2 + |b_i|^2))$ and $b/(1 + \alpha_i(|\vec{w}_i|^2 + |b_i|^2))$ provide normalizations for the weights and biases. Namely, if we fix $\alpha_i > 0$, the norm of $\frac{\vec{w}_i \cdot v_{\text{in}} + b_i}{1+\alpha_i(|\vec{w}_i|^2+|b_i|^2)}$ is bounded for arbitrary $\vec{w}_i$, b_i and bounded v_{in}. This means even if the gradient for $(\vec{w}_i, b_i)$, and hence the updated $(\vec{w}_i, b_i)$, is large during the learning process, the above operation is still bounded (even if an unbounded function σ_i is used, such as ReLu).

The following computer code implements (26) in Python using TENSORFLOW. It appears in the definition of call for the dense layer that will be followed by an activation layer. A similar change can also be made to a convolution layer.

```
norm_w = tf.math.reduce_sum(tf.math.square(self.w),axis=0,keepdims=True)
norm_b = tf.math.square(self.b)
H = tf.math.reciprocal(1 + self.alpha * (norm_w + norm_b))
y = tf.matmul(inputs, self.w) + self.b
return H * y
```

Applying Theorem 4.15, we have the metric

$$\begin{aligned} H_T(\vec{\alpha}) = \sum_i tr\,(&H_i(\alpha_i) \cdot \partial\rho_i \cdot \mathscr{I}_i \cdot (\partial\rho_i)^* \\ &- H_i(\alpha_i) \cdot \rho_i \cdot \mathscr{I}_i \cdot (\partial\rho_i)^* \cdot H_i(\alpha_i) \cdot (\partial\rho_i) \cdot \mathscr{I}_i \cdot \rho_i^*) \end{aligned}$$

that can be used in the gradient descent, where $\mathscr{I}_i$ denotes the matrix $\begin{pmatrix} I_{d_i} & 0 \\ 0 & \text{diag}(\alpha_\gamma) \end{pmatrix}$ and $\text{diag}(\alpha_\gamma)$ denotes the diagonal matrix with entries α_γ (see also (23)). With the simplifying assumptions above as in (25), $H_T(\vec{\alpha}) = \bigoplus_i (H_T)_i(\alpha_i)$ is equal to

$$\begin{aligned} (H_T)_i(\alpha_i) &= H_i(\alpha_i)\left(I_{d_i+1} - (\alpha_i H_i(\alpha_i))\tilde{w}_i^* \cdot \tilde{w}_i\right) \\ &= H_i(\alpha_i)\left(I_{d_i+1} - \frac{\tilde{w}_i^* \cdot \tilde{w}_i}{\alpha_i^{-1} + |\tilde{w}_i|^2}\right) > 0 \end{aligned}$$

where $\tilde{w}_i$ denotes the row vector $(\vec{w}_i\ b_i)$. We remark that it is valid to use different parameters $\vec{\alpha}$ and $\vec{\alpha}'$ for the bundle metric $H_i(\alpha_i)$ and the Riemannian metric $H_T(\vec{\alpha}')$ respectively, if we take $\vec{\alpha} > 0$ and $\vec{\alpha}' \geq 0$. (Here a vector is said to be > 0 or ≥ 0 if each of its entries is). For instance, the bundle metric $H_i(\alpha_i)$ with $\alpha_i > 0$ is well-defined for $\mathcal{V}_i$ over the Euclidean space $\mathcal{M}^0$.

Composing the operations (26) at different neurons allows us to obtain a machine function $f_{\tilde{w}}$. This composition is known as the forward propagation. To do machine learning, we minimize various quantities concerning $f_{\tilde{w}}$, for instance the distance

$$E(\tilde{w}) = d(f_{\tilde{w}}, f)$$

with a given function f. ($\tilde{w}$ denotes the tuples of all $\tilde{w}_i$ for all vertices i.)

In the Riemannian metric $H_T(\vec{\alpha})$, the gradient descent of E over $\mathcal{M}$ is given by the matrix multiplication

$$-(\partial_{\tilde{w}_i} E) \cdot ((H_T)_i(\alpha_i))^{-1}$$

where the differential $\partial_{\tilde{w}_i} E = (\partial_{\vec{w}_i} E\ \partial_{b_i} E)$ is a row vector. The inverse matrix $(I_{d_i} - (\alpha_i H_i(\alpha_i))\tilde{w}_i^* \cdot \tilde{w}_i)^{-1}$ can be approximated by $I_{d_i} + (\alpha_i H_i(\alpha_i))\tilde{w}_i^* \cdot \tilde{w}_i$. Thus, the gradient descent can be approximated by

$$-H_i(\alpha_i)^{-1}\, \partial_{\tilde{w}_i} E - \alpha_i(\partial_{\tilde{w}_i} E \cdot \tilde{w}_i^*)\tilde{w}_i \tag{27}$$

which is the update to the current weights $\tilde{w}_i$. The partial derivatives $\partial_{\tilde{w}_i} E$ can be efficiently computed by the chain rule for the composition function E, which is well-known as the backward propagation algorithm. Note that for $\alpha_i = 0$, (27) reduces back to the Euclidean gradient descent $-\partial_{\tilde{w}_i} E$.

The following code implements the gradient descent (27) in the definition of train_step in Keras Model class. It is placed after taking gradient tape of TENSORFLOW which records the partial derivatives by grads[i]. Let the i-th and $(i+1)$-th trainable variables be weights and biases.

```
w = trainable_vars[i]
b = trainable_vars[i+1]
norm_w = tf.math.reduce_sum(tf.math.square(w),axis=0)
norm_b = tf.math.square(b)
H_inverse = 1 + alpha * (norm_b + norm_w)
dotprod = tf.math.reduce_sum(tf.multiply(grads[i],w),axis=0)+\
          tf.multiply(grads[i+1],b)
grads[i] = H_inverse * grads[i] + alpha * tf.multiply(dotprod,w)
grads[i+1] = H_inverse * grads[i+1] + alpha * tf.multiply(dotprod,b)
```

As explained above, we expect from the theory side that the bundle metric $H_i(\alpha_i)$ can be used as a weight normalization in (26) to handle gradient explosion, with the parameters α_i encoding the significance of the i-neuron. Moreover, the gradient descent over a compact moduli $\mathcal{M}$ with the global metric $H_T(\vec{\alpha})$ for $\vec{\alpha} > 0$ must converge mathematically, although we may need to take a change of coordinates and perform the gradient descent in other charts of $\mathcal{M}$ as well. We will leave further experiments as a part of future work.

Acknowledgment

We are grateful to Bernd Henschenmacher for informing us of the work [22] which adds very interesting historical context and motivation for the application of the techniques in Section 3 towards quantum mechanics.

References

[1] M. Armenta, T. Brüstle, S. Hassoun, and M. Reineke. Double framed moduli spaces of quiver representations. *Linear Algebra and its Applications*, **650**:98–131, 2022. MR4439387

[2] M. A. Armenta and P.-M. Jodoin. The representation theory of neural networks. *preprint*, 2020. arXiv:2007.12213.

[3] Y. Berest, G. Khachatryan, and A. Ramadoss. Derived representation schemes and cyclic homology. *Adv. Math.*, **245**:625–689, 2013. MR3084440

[4] G. Birkhoff. On the structure of abstract algebras. *Mathematical Proceedings of the Cambridge Philosophical Society*, **31**(4):433–454, 1935.

[5] Y. Chen, Y. Huang, and N. Leung. Embeddings from noncompact symmetric spaces to their compact duals.

[6] M.C.N. Cheng, V. Anagiannis, M. Weiler, P. de Haan, T.S. Cohen, and M. Welling. Covariance in physics and convolutional neural networks. *preprint*, 2019. arXiv:1906.02481.

[7] T.S. Cohen, M. Geiger, J. Koehler, and M. Welling. Spherical CNNs. *ICLR*, 2018.

[8] T.S. Cohen, M. Geiger, and M. Weiler. A general theory of equivariant cnns on homogeneous spaces. *NeurIPS*, 2019. arXiv:1811.02017.

[9] T.S. Cohen, M. Weiler, B. Kicanaoglu, and M. Welling. Gauge equivariant convolutional networks and the icosahedral CNN. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2019.

[10] T.S. Cohen and M. Welling. Group equivariant convolutional networks. In *Proceedings of The 33rd International Conference on Machine Learning*, volume 48, pages 2990–2999, 2016.

[11] A. Connes. Noncommutative differential geometry. *Inst. Hautes Études Sci. Publ. Math.*, (**62**):257–360, 1985. MR0823176

[12] J. Cuntz and D. Quillen. Algebra extensions and nonsingularity. *J. Amer. Math. Soc.*, **8**(2):251–289, 1995. MR1303029

[13] P. de Haan, T. Cohen, and M. Welling. Natural graph networks. *preprint*, 2020. arXiv:2007.08349.

[14] S. Eilenberg and B. Tilson. *Automata, Languages, and Machines.* ISSN. Elsevier Science, 1976. MR0530383

[15] O.-E. Ganea, G. Bécigneul, and T. Hofmann. Hyperbolic entailment cones for learning hierarchical embeddings. *ArXiv*, abs/1804.01882, 2018.

[16] O.-E. Ganea, G. Bécigneul, and T. Hofmann. Hyperbolic neural networks. *ArXiv*, abs/1805.09112, 2018.

[17] I. Ganev and R. Walters. The QR decomposition for radial neural networks. *preprint*, 2021. arXiv:2107.02550 .

[18] V. Ginzburg. Lectures on noncommutative geometry. *preprint*, 2005. arXiv:0506603.

[19] G. Jeffreys and S.-C. Lau. Kähler geometry of quiver varieties and machine learning. *preprint*, 2021. arXiv:2101.11487.

[20] A.D. King. Moduli of representations of finite-dimensional algebras. *Quart. J. Math. Oxford Ser. (2)*, **45**(180):515–530, 1994. MR1315461

[21] S. C. Kleene. *Representation of Events in Nerve Nets and Finite Automata*, pages 3–42. Princeton University Press, 2016. MR0077478

[22] M. Liebmann, H. Ruhaak, and B. Henschenmacher. Nonassociative algebras and quantum physics – a historical perspective. *preprint*, 2019. arXiv:1909.04027.

[23] H. Nakajima. Varieties associated with quivers. In *Representation theory of algebras and related topics (Mexico City, 1994)*, volume 19 of *CMS Conf. Proc.*, pages 139–157. Amer. Math. Soc., Providence, RI, 1996. MR1388562

[24] H. Nakajima. Quiver varieties and finite-dimensional representations of quantum affine algebras. *J. Amer. Math. Soc.*, **14**(1):145–238, 2001. MR1808477

[25] M. Nickel and D. Kiela. Poincaré embeddings for learning hierarchical representations. In *NIPS*, 2017.

[26] M. Reineke. Framed quiver moduli, cohomology, and quantum groups. *J. Algebra*, **320**(1):94–115, 2008. MR2417980

[27] J. Reiterman. The Birkhoff theorem for finite algebras. *Algebra Universalis*, 14:1–10, 12 1982. MR0634411

[28] F. Sala, C. De Sa, A. Gu, and C. Ré. Representation tradeoffs for hyperbolic embeddings. *Proceedings of Machine Learning Research*, **80**:4460–4469, 2018.

[29] M. P. Schützenberger. Une théorie algébrique du codage. *Séminaire Dubreil. Algèbre et théorie des nombres*, **9**:1–24, 1955-1956.

[30] M.P. Schützenberger. On finite monoids having only trivial subgroups. *Information and Control*, **8**(2):190–194, 1965. MR0176883

[31] A. Sheshmani and Y. You. Categorical representation learning: Morphism is all you need. *preprint*, 2021. arXiv:2103.14770.

[32] A. Tacchella. An introduction to associative geometry with applications to integrable systems. *J. Geom. Phys.*, **118**:202–233, 2017. MR3660912

George Jeffreys
Department of Mathematics and Statistics
Boston University
111 Cummington Mall
Boston, MA 02215
USA
E-mail: georgej@bu.edu

Siu-Cheong Lau
Department of Mathematics and Statistics
Boston University
111 Cummington Mall
Boston, MA 02215
USA
E-mail: lau@math.bu.edu