Mathematical Theory of Recursive Symbol Systems

Peter Fletcher
E-mail: maa03@keele.ac.uk

Technical Report GEN-0401

Department of Computer Science
School of Computing and Mathematics
Keele University
Keele, Staffs, ST5 5BG
U.K.

Tel: +441782 583260
Fax: +441782 713082

November 2004

Abstract

This paper introduces a new type of graph grammar for representing recursively structured two-dimensional geometric patterns, for the purposes of syntactic pattern recognition. The grammars are noise-tolerant, in the sense that they can accommodate patterns that contain pixel noise, variability in the geometric relations between symbols, missing parts, or ambiguous parts, and that overlap other patterns. Geometric variability is modelled using *fleximaps*, drawing on concepts from the theory of Lie algebras and tensor calculus.

The formalism introduced here is intended to be generalisable to all problem domains in which complex, many-layered structures occur in the presence of noise.

Keywords: graph grammars, context-free grammars, recursive symbol processing, parsing, syntactic pattern recognition, noise tolerance, affine invariance.

Table of Contents

§1. Introduction 4

§2. Algebraic and analytic preliminaries 7

§3. Affine transformations 16

§4. Matching a template to an image 30

§5. Fleximaps 33

§6. Networks and homomorphisms 36

§7. The recognition problem 39

References 47

1. Introduction

1.1 Aims

This work is an exercise in syntactic pattern recognition. The aim is to represent complex geometric patterns as symbol systems, in a robust way that can cope with missing or deformed symbols, variability in the geometric relations between parts of a symbol, and other types of noise. More generally, the long-term aim of my research is to combine the expressive power of recursive symbol systems (such as formal grammars) with the noise-tolerance of massively parallel, distributed computational systems (such as neural networks).

A *symbol system* is a structure consisting of finitely many elements of various types, called *symbols*; a symbol may *contain* other symbols, and may stand in certain relations to *neighbouring* symbols. The permitted relations of containment and neighbourhood are specified by a *grammar*, which is also a symbol system. The grammar is finite, but by means of recursion it can describe infinitely many symbol systems. The most familiar examples of grammars are those used to describe programming languages and natural languages; these are called *string* grammars, because they describe *sequences* of symbols. *Graph* grammars are more powerful than string grammars, as they can represent more complex, non-sequential configurations of symbols.

The importance of recursive symbol systems was emphasised in the so-called 'classical' tradition of artificial intelligence and cognitive science; it became the subject of controversy between the 'classical' tradition and connectionism in the 1980s and 1990s (Smolensky, 1988; Fodor & Pylyshyn, 1988; Barnden & Pollack, 1991; Dinsmore, 1992; Horgan & Tienson, 1996; Garfield, 1997; Sougné, 1998; Hadley, 1999). There is no doubt that recursive symbol systems have an expressive power and computational flexibility that is unmatched by any other type of representation. Their big weakness is their 'brittleness': their inability to cope with noise, vagueness, incomplete or contradictory information, ambiguity, context-dependency, and so on. The aim of my work is to overcome this brittleness.

In this paper I shall set up a mathematical theory of noisy symbol systems and of how they may be used to represent complex, recursively structured two-dimensional geometric patterns. (My choice of geometric patterns as a problem domain is simply for the sake of convenience; the theory is intended to be of wider application.)

The grammars I shall use are essentially a kind of graph grammar. In my previous work (Fletcher, 2001) I showed how regular graph grammars describing two-dimensional geometric patterns could be parsed and learned from positive examples. The present paper goes beyond this in the following ways.

- The grammatical framework is extended from *regular* to *context-free* graph grammars.
- The full geometry of the plane is used, whereas previously the plane was represented by a discrete square grid.
- Noise is permitted, in the form of pixel noise, variability in the geometric relations between parts, incomplete patterns, and overlapping patterns.

1.2 Terminology

We must distinguish between symbol types and symbol tokens. The Roman alphabet consists of 26 letters, in upper and lower case, which gives 52 symbol types. A particular occurrence of a symbol type at a particular position, orientation and size is a symbol token. Thus a text document may contain thousands of symbol *tokens*, each of which is an instance of one of the 52 symbol *types* (ignoring punctuation marks and other non-letter symbols).

A *grammar* is a system of symbol types. A *pattern* is the system of all symbol tokens present in a given image. A grammar specifies a (usually infinite) set of possible patterns.

The shape of a symbol type is described by a *template*, which depicts how a token of that type would look, occurring at a standard position, orientation and size, without noise. Each symbol type has a template.

The position, orientation, size, and degree of stretching and shearing of a symbol token in an image is specified by an affine transformation, known as the *embedding* of the token. It is a mapping from the template of the symbol token's type into the image plane.

1.3 The contents of this paper

§2 provides necessary algebraic and analytic preliminaries from the theory of real finite-dimensional vector spaces and linear transformations. All the material here is standard, but to ensure that this paper is self-contained and rigorous I have re-derived all results in precisely the form in which I shall need them in the later sections.

§3 sets up the basic theory of affine transformations, which are used to describe the embedding of symbol tokens in the image plane and the relations of symbol tokens to one another. I use coordinate-free concepts from the theory of Lie algebras and tensor calculus but also give coordinate-based calculation techniques for use in a computational implementation. (See Price (1977), Hausner & Schwartz (1968) and Boothby (1986) for more details on these mathematical concepts.)

§4 deals with templates, which describe the concrete appearance of symbols in the image plane. The goodness of match of a template with an image is measured by a correlation function; I calculate the derivative of this function so that it can be maximised.

§5 defines the concept of a *fleximap*, which is used to represent a flexible affine transformation, i.e., one that can be deformed from its nominal value along certain degrees of freedom.

§6 defines the central concept of a *network*, which is used to represent symbol systems (both grammars and patterns). The parsing of a pattern under a grammar is represented by a *homomorphism* between networks (rather than by a parse tree, as in conventional grammatical formalisms).

§7 sets up the problem of recognising the pattern in a given image. I state the recognition problem formally with the help of a *Match* function and show how the embedding of the pattern can be optimised by gradient ascent on the *Match* function.

The next step (to be dealt with in a future paper) is to introduce a parsing algorithm to find the best pattern and homomorphism. The problem of learning the grammar from a given set of example images is left for future work.

In this paper lemmas and theorems will be numbered consecutively $1, 2, 3, \ldots$ within each section, and will be cited outside the section by adding the section number: e.g., lemma 2.3 is lemma 3 of §2.

2. Algebraic and Analytic Preliminaries

2.1 The norm of a vector

Let V be a real finite-dimensional vector space. Choose a basis $\{e_1, \ldots e_n\}$, and define a norm on V by

$$\left|\sum_{i=1}^{n} v^i e_i\right| = \sqrt{\sum_{i=1}^{n}(v^i)^2}.$$

This norm is said to be *induced* by the basis.

LEMMA 1. The norm satisfies the usual axioms:

(i) $\forall u, v \in V \ |u+v| \leq |u| + |v|$

(ii) $\forall k \in \mathbf{R} \forall v \in V \ |kv| = |k||v|$

(iii) $v \neq 0 \Rightarrow |v| > 0$

Proof. (i) For any $u = \sum_{i=1}^{n} u^i e_i$ and any $v = \sum_{i=1}^{n} v^i e_i$, by the Cauchy-Schwarz inequality we have

$$\begin{aligned} |u+v|^2 &= \sum_{i=1}^{n}(u^i + v^i)^2 = \sum_{i=1}^{n}(u^i)^2 + \sum_{i=1}^{n}(v^i)^2 + 2\sum_{i=1}^{n} u^i v^i \\ &\leq \sum_{i=1}^{n}(u^i)^2 + \sum_{i=1}^{n}(v^i)^2 + 2\sqrt{\sum_{i=1}^{n}(u^i)^2}\sqrt{\sum_{i=1}^{n}(v^i)^2} = |u|^2 + |v|^2 + 2|u||v| = (|u| + |v|)^2 \end{aligned}$$

so $|u+v| \leq |u| + |v|$.

(ii) For any $k \in \mathbf{R}$ and any $v = \sum_{i=1}^{n} v^i e_i$,

$$|kv|^2 = \sum_{i=1}^{n}(kv^i)^2 = k^2 \sum_{i=1}^{n}(v^i)^2 = k^2|v|^2$$

so $|kv| = |k||v|$.

(iii) For any $v = \sum_{i=1}^{n} v^i e_i$, if $v \neq 0$ then we must have at least one $v^i \neq 0$, so $|v|^2 = \sum_{i=1}^{n}(v^i)^2 > 0$. ∎

LEMMA 2. If a norm $|\cdot|$ is induced by a basis $\{e_1, \ldots e_n\}$, and a second norm $|\cdot|'$ is induced by another basis $\{f_1, \ldots f_n\}$, then the two norms are related by

$$\exists \alpha \in \mathbf{R} \ \forall v \in V \ |v| \leq \alpha |v|', \quad \exists \beta \in \mathbf{R} \ \forall v \in V \ |v|' \leq \beta |v|.$$

Proof. Since $\{e_1, \ldots e_n\}$ is a basis, every vector f_j can be expressed as a linear combination $f_j = \sum_{i=1}^{n} M^i{}_j e_i$, for some real numbers $M^i{}_j$.

Any vector $v \in V$ can be expressed in terms of the basis $\{f_1, \ldots f_n\}$ as $\sum_{j=1}^{n} b^j f_j$. Then

$$v = \sum_{j=1}^{n} b^j f_j = \sum_{j=1}^{n}\sum_{i=1}^{n} b^j M^i{}_j e_i = \sum_{i=1}^{n}\Big(\sum_{j=1}^{n} b^j M^i{}_j\Big) e_i$$

so, by the definition of the norms and the Cauchy-Schwarz inequality,

$$|v|^2 = \sum_{i=1}^{n}\Big(\sum_{j=1}^{n} b^j M^i{}_j\Big)^2 \leq \sum_{i=1}^{n}\Big(\sum_{j=1}^{n}(b^j)^2\Big)\Big(\sum_{j=1}^{n}(M^i{}_j)^2\Big) = |v|'^2 \sum_{i,j=1}^{n}(M^i{}_j)^2$$

so taking $\alpha = \sqrt{\sum_{i,j=1}^{n}(M^i{}_j)^2}$ verifies the first half of the lemma. The other half is proved similarly. ∎

2.2 Convergence

The norm induced by a basis induces a topology on V:

$$x_k \to l \text{ as } k \to \infty \quad \text{iff} \quad \forall \varepsilon > 0\, \exists N\, \forall k > N\ |x_k - l| < \varepsilon.$$

By lemma 2, the topology is independent of the choice of basis.

2.3 Vanishing and limited functions

Let V and W be real finite-dimensional vector spaces, each with a norm induced by a basis.

- A function $f\colon V \to W$ is *vanishing* iff $\forall \varepsilon > 0\, \exists \delta > 0\, \forall v \in V\ |v| < \delta \ \Rightarrow\ |f(v)| \le \varepsilon |v|$.
- A function $f\colon V \to W$ is *limited* iff $\exists \varepsilon > 0\, \exists \delta > 0\, \forall v \in V\ |v| < \delta \ \Rightarrow\ |f(v)| \le \varepsilon |v|$.

By lemma 2, these concepts are independent of the choice of bases. In the lemmas that follow, U, V and W are any real finite-dimensional vector spaces, each with a norm induced by a basis.

LEMMA 3. Any vanishing function is limited.

Proof. Immediate from the definitions. ∎

LEMMA 4. Any linear function is limited and continuous.

Proof. Let $F\colon V \to W$ be linear. Let $\{e_1, \dots e_m\}$ be the basis on V and $\{f_1, \dots f_n\}$ the basis on W used to induce the norms. Let A be the $n \times m$ matrix representing F, i.e., $F(e_i) = \sum_{j=1}^n A^j{}_i f_j$. Choose $\varepsilon = \sqrt{\sum_{i=1}^m \sum_{j=1}^n (A^j{}_i)^2}$. Then, for any $v = \sum_{i=1}^m v^i e_i$ in V, $F(v) = \sum_{j=1}^n \left(\sum_{i=1}^m A^j{}_i v^i\right) f_j$, and, by the Cauchy-Schwarz inequality,

$$|F(v)|^2 = \sum_{j=1}^n \left(\sum_{i=1}^m A^j{}_i v^i\right)^2 \le \sum_{j=1}^n \left(\sum_{i=1}^m (A^j{}_i)^2\right) \left(\sum_{i=1}^m (v^i)^2\right) = \varepsilon^2 |v|^2$$

so $|F(v)| \le \varepsilon |v|$. This verifies that F is limited (δ may be chosen arbitrarily), and also implies that F is continuous, since

$$\forall x, a \in V \quad |F(x) - F(a)| = |F(x - a)| \le \varepsilon |x - a|.$$

∎

LEMMA 5. If $f\colon V \to W$ is limited then $f(0) = 0$ and f is continuous at 0.

Proof. There exist positive ε_0, δ_0 such that

$$\forall v \in V \quad |v| < \delta_0 \ \Rightarrow\ |f(v)| \le \varepsilon_0 |v|.$$

Taking $v = 0$ gives $|f(0)| \le \varepsilon_0 |0| = 0$, so $f(0) = 0$ as required.

To show that f is continuous at 0, for any $\varepsilon > 0$, choose $\delta = \min(\delta_0, \frac{\varepsilon}{\varepsilon_0})$; then

$$\forall v \quad |v| < \delta \ \Rightarrow\ |f(v)| \le \varepsilon_0 |v| < \varepsilon_0 \delta \le \varepsilon$$

as required. ∎

LEMMA 6. If $f,g\colon V \to W$ are vanishing and $k \in \mathbf{R}$ then $f+g$ and kf are also vanishing.

Proof. For any $\varepsilon > 0$, we know that there exist $\delta_1 > 0$ and $\delta_2 > 0$ such that

$$\forall v \quad |v| < \delta_1 \;\Rightarrow\; |f(v)| \le \frac{\varepsilon}{2}|v|,$$
$$\forall v \quad |v| < \delta_2 \;\Rightarrow\; |g(v)| \le \frac{\varepsilon}{2}|v|.$$

Now, choose $\delta = \min(\delta_1, \delta_2)$. Then

$$\forall v \quad |v| < \delta \;\Rightarrow\; |(f+g)(v)| \le |f(v)| + |g(v)| \le \varepsilon|v|$$

so $f+g$ is vanishing.

Secondly, for any $\varepsilon > 0$, we know that there exists $\delta > 0$ such that

$$\forall v \quad |v| < \delta \;\Rightarrow\; |f(v)| \le \frac{\varepsilon}{|k|+1}|v|.$$

Hence

$$\forall v \quad |v| < \delta \;\Rightarrow\; |(kf)(v)| = |k||f(v)| \le \varepsilon\frac{|k|}{|k|+1}|v| < \varepsilon|v|$$

so kf is vanishing. ∎

LEMMA 7. If $f,g\colon V \to W$ are limited and $k \in \mathbf{R}$ then $f+g$ and kf are also limited.

Proof. We know that there exist positive $\varepsilon_1, \delta_1, \varepsilon_2, \delta_2$ such that

$$\forall v \quad |v| < \delta_1 \;\Rightarrow\; |f(v)| \le \varepsilon_1|v|,$$
$$\forall v \quad |v| < \delta_2 \;\Rightarrow\; |g(v)| \le \varepsilon_2|v|.$$

Now, choose $\varepsilon = \varepsilon_1 + \varepsilon_2$ and $\delta = \min(\delta_1, \delta_2)$. Then

$$\forall v \quad |v| < \delta \;\Rightarrow\; |(f+g)(v)| \le |f(v)| + |g(v)| \le \varepsilon|v|$$

so $f+g$ is limited.

For the second part, choose $\varepsilon = (|k|+1)\varepsilon_1 > 0$. Then

$$\forall v \quad |v| < \delta_1 \;\Rightarrow\; |(kf)(v)| = |k||f(v)| \le \varepsilon|v|$$

so kf is limited. ∎

LEMMA 8. If $f\colon V \to W$ is vanishing and $g\colon U \to V$ is limited then $f \circ g$ is vanishing.

Proof. We know that there exist positive ε_1, δ_1 such that

$$\forall u \in U \quad |u| < \delta_1 \;\Rightarrow\; |g(u)| \le \varepsilon_1|u|.$$

Also, for any $\varepsilon > 0$, we know that there exists $\delta_2 > 0$ such that

$$\forall v \in V \quad |v| < \delta_2 \;\Rightarrow\; |f(v)| \le \frac{\varepsilon}{\varepsilon_1}|v|.$$

Choose $\delta = \min(\delta_1, \frac{\delta_2}{\varepsilon_1})$. Then

$$\forall u \quad |u| < \delta \;\Rightarrow\; |g(u)| \le \varepsilon_1|u| < \varepsilon_1\delta \le \delta_2 \;\Rightarrow\; |f(g(u))| \le \frac{\varepsilon}{\varepsilon_1}|g(u)| \le \varepsilon|u|$$

as required. ∎

LEMMA 9. If $f\colon V\to W$ is limited and $g\colon U\to V$ is vanishing then $f\circ g$ is vanishing.

Proof. We know that there exist positive ε_2, δ_2 such that

$$\forall v\in V \quad |v|<\delta_2 \;\Rightarrow\; |f(v)|\le \varepsilon_2|v|.$$

Also, for any $\varepsilon>0$, we know that there exists $\delta_1>0$ such that

$$\forall u\in U \quad |u|<\delta_1 \;\Rightarrow\; |g(u)|\le \min\bigl(\frac{\varepsilon}{\varepsilon_2},\delta_2\bigr)|u|.$$

Choose $\delta=\min(\delta_1,1)$. Then, for any u with $|u|<\delta$, we have $|u|<\delta_1$, so $|g(u)|\le \min\bigl(\frac{\varepsilon}{\varepsilon_2},\delta_2\bigr)|u|$, so $|g(u)|\le\delta_2|u|<\delta_2$ (since $|u|<1$). Hence $|f(g(u))|\le\varepsilon_2|g(u)|$. But we also have $|g(u)|\le\frac{\varepsilon}{\varepsilon_2}|u|$, so $|f(g(u))|\le\varepsilon|u|$ as required. ∎

2.4 Application of the theory to linear transformations

For any real finite-dimensional vector spaces V and W, let $\mathcal{L}(V,W)$ be the vector space of all linear transformations from V to W. A basis $\{e_1,\dots e_m\}$ for V and a basis $\{f_1,\dots f_n\}$ for W induce a basis $\{\,h^i{}_j \mid 1\le i\le m, 1\le j\le n\,\}$ for $\mathcal{L}(V,W)$, where

$$h^i{}_j(e_k)=\begin{cases} f_j & \text{if } i=k\\ 0 & \text{if } i\ne k\end{cases}$$

Then any $A\in\mathcal{L}(V,W)$ can be represented in terms of this basis by

$$A=\sum_{i=1}^{m}\sum_{j=1}^{n}A^j{}_i h^i{}_j$$

giving

$$A(e_i)=\sum_{j=1}^{n}A^j{}_i f_j.$$

For any element $v=\sum_{i=1}^m v^i e_i$ of V we have

$$A(v)=\sum_{j=1}^{n}\Bigl(\sum_{i=1}^{m}A^j{}_i v^i\Bigr)f_j.$$

Assume that V, W and $\mathcal{L}(V,W)$ have the norms induced by these bases. For any $A=\sum_{i=1}^m\sum_{j=1}^n A^j{}_i h^i{}_j$ we have

$$|A|^2=\sum_{i=1}^{m}\sum_{j=1}^{n}(A^j{}_i)^2=\sum_{i=1}^{m}|A(e_i)|^2.$$

In the lemmas that follow, let U,V,W be any real finite-dimensional vector spaces with norms induced by bases.

LEMMA 10. If $v\in V$ and $A\in\mathcal{L}(V,W)$ then $|Av|\le|A||v|$.

Proof. Let $v=\sum_{i=1}^m v^i e_i$. Then

$$|A(v)|^2=\Bigl|\sum_{i=1}^{m}v^iA(e_i)\Bigr|^2\le\Bigl(\sum_{i=1}^{m}|v^i||A(e_i)|\Bigr)^2\le\sum_{i=1}^{m}(v^i)^2\sum_{i=1}^{m}|A(e_i)|^2=|v|^2|A|^2$$

using the triangle inequality and the Cauchy-Schwarz inequality. Hence $|A(v)|\le|A||v|$. ∎

LEMMA 11. If $A \in \mathcal{L}(V,W)$ and $B \in \mathcal{L}(U,V)$ then $|AB| \leq |A||B|$.

Proof. Let $\{d_1, \ldots d_l\}$ be the basis on U that induces the norm. By the previous lemma,

$$|AB|^2 = \sum_{k=1}^{l} |AB(d_k)|^2 \leq \sum_{k=1}^{l} |A|^2|B(d_k)|^2 = |A|^2|B|^2$$

so $|AB| \leq |A||B|$. ▮

LEMMA 12. If (A_k) is a sequence in $\mathcal{L}(V,W)$ and $L \in \mathcal{L}(V,W)$ then

$$A_k \to L \text{ as } k \to \infty \quad \text{iff} \quad \forall v \in V\ A_k(v) \to L(v) \text{ as } k \to \infty.$$

(A special case of this is: $\sum_{k=1}^{\infty} V_k = L$ iff $\forall v \in V \sum_{k=1}^{\infty} V_k(v) = L(v)$.)

Proof. If $A_k \to L$ as $k \to \infty$ then, for any $v \in V$,

$$|A_k(v) - L(v)| = |(A_k - L)(v)| \leq |A_k - L||v|$$

so, by comparison, $A_k(v) \to L(v)$ as $k \to \infty$, as required.

Conversely, suppose $\forall v\ A_k(v) \to L(v)$ as $k \to \infty$. Let $\{e_1, \ldots e_m\}$ be the basis that induces the norm on V. Then

$$|A_k - L|^2 = \sum_{i=1}^{m} |(A_k - L)(e_i)|^2 \to 0 \text{ as } k \to \infty$$

as required. ▮

For any normed vector space S and any $\varepsilon > 0$, let $B_\varepsilon(S) = \{\, v \in S \mid |v| < \varepsilon \,\}$.

LEMMA 13. If (c_n) is a real sequence, and $\sum_{n=0}^{\infty} c_n l^n$ is absolutely convergent for some positive real l, then $\sum_{n=0}^{\infty} c_n A^n$ is absolutely convergent for all $A \in B_l(\mathcal{L}(V,V))$, and if f is a function satisfying

$$\forall A \in B_l(\mathcal{L}(V,V)) \quad f(A) = \sum_{n=0}^{\infty} c_n A^n$$

then f is continuous on $B_l(\mathcal{L}(V,V))$.

Proof. Consider any $A \in B_l(\mathcal{L}(V,V))$. Let $a = |A| < l$, $\theta = \frac{l+a}{2}$, and $\varepsilon = \frac{l-a}{2}$. For all n, we have $|c_n A^n| \leq |c_n a^n| \leq |c_n l^n|$, so $\sum_{n=0}^{\infty} c_n A^n$ is absolutely convergent by comparison with $\sum_{n=0}^{\infty} c_n l^n$.

Also, for any $X \in B_\varepsilon(\mathcal{L}(V,V))$, we have

$$|(A+X)^n - A^n| = |R_n(A,X)| \leq R_n(a,x)$$

where $x = |X|$, $R_n(A,X)$ is the sum of the $2^n - 1$ terms in the expansion of $(A+X)^n$ other than A^n, and $R_n(a,x)$ is the same sum of terms with A and X replaced by the real numbers a and x.

Now,

$$\begin{aligned}
0 \le R_n(a,x) &= (a+x)^n - a^n = x\big((a+x)^{n-1} + (a+x)^{n-2}a + \cdots + (a+x)a^{n-2} + a^{n-1}\big) \\
&\le x(n\theta^{n-1}) \quad \text{since } a < \theta \text{ and } a+x < \theta \\
&\le x\frac{(\theta+\varepsilon)^n}{\varepsilon} \quad \text{since } (\theta+\varepsilon)^n = \theta^n + n\theta^{n-1}\varepsilon + \cdots + \varepsilon^n \\
&= \frac{x}{\varepsilon} l^n.
\end{aligned}$$

Hence $\sum_{n=0}^{\infty} |c_n R_n(a,x)|$ exists, by comparison with $\sum_{n=0}^{\infty} |c_n l^n|$, with

$$\sum_{n=0}^{\infty} |c_n R_n(a,x)| \le \frac{x}{\varepsilon} \sum_{n=0}^{\infty} |c_n l^n| = Mx$$

where $M = \frac{1}{\varepsilon}\sum_{n=0}^{\infty} |c_n l^n|$. Hence, by comparison, $\sum_{n=0}^{\infty} |c_n(A+X)^n - c_n A^n|$ also exists, with

$$\sum_{n=0}^{\infty} |c_n(A+X)^n - c_n A^n| \le \sum_{n=0}^{\infty} |c_n R_n(a,x)| \le Mx.$$

Now, we have $A, A+X \in B_l(\mathcal{L}(V,V))$, so $f(A) = \sum_{n=0}^{\infty} c_n A^n$ and $f(A+X) = \sum_{n=0}^{\infty} c_n (A+X)^n$, so

$$|f(A+X) - f(A)| = \left|\sum_{n=0}^{\infty} c_n(A+X)^n - c_n A^n\right| \le \sum_{n=0}^{\infty} |c_n(A+X)^n - c_n A^n| \le Mx.$$

Hence f is continuous at A, as required. ∎

LEMMA 14. If (c_n) is a real sequence, and $\sum_{n=2}^{\infty} c_n l^n$ is absolutely convergent for some positive real l, and $f\colon \mathcal{L}(V,V) \to \mathcal{L}(V,V)$ satisfies

$$\forall A \in B_l(\mathcal{L}(V,V)) \quad f(A) = \sum_{n=2}^{\infty} c_n A^n,$$

then f is vanishing.

Proof. For any A in $B_l(\mathcal{L}(V,V))$ and any $n \ge 2$, we have

$$|c_n A^n| \le |c_n||A|^n \le \frac{|A|^2}{l^2}|c_n l^n|$$

so $\sum_{n=2}^{\infty} |c_n A^n|$ converges, by comparison with $\sum_{n=2}^{\infty} |c_n l^n|$, and

$$|f(A)| = \left|\sum_{n=2}^{\infty} c_n A^n\right| \le \sum_{n=2}^{\infty} |c_n A^n| \le \sum_{n=2}^{\infty} \frac{|A|^2}{l^2}|c_n l^n| = M|A|^2$$

where $M = \frac{1}{l^2}\sum_{n=2}^{\infty} |c_n l^n|$. Hence we can show that f is vanishing: for any $\varepsilon > 0$, choose $\delta = \min(l, \frac{\varepsilon}{M+1})$, giving

$$\forall A \quad |A| < \delta \;\Rightarrow\; |f(A)| \le M|A|^2 \le M\delta|A| \le \varepsilon|A|$$

as required. ∎

LEMMA 15. If $f, g: V \to \mathcal{L}(W,W)$ are limited functions, and $h: V \to \mathcal{L}(W,W)$ is defined by $\forall v \in V\ h(v) = f(v)g(v)$, then h is vanishing.

Proof. We know that there exist positive $\varepsilon_1, \delta_1, \varepsilon_2, \delta_2$ such that

$$\begin{aligned}\forall v \quad |v| < \delta_1 &\Rightarrow |f(v)| \le \varepsilon_1 |v|,\\ \forall v \quad |v| < \delta_2 &\Rightarrow |g(v)| \le \varepsilon_2 |v|.\end{aligned}$$

Now, for any $\varepsilon > 0$, choose $\delta = \min(\delta_1, \delta_2, \frac{\varepsilon}{\varepsilon_1\varepsilon_2})$. Then

$$\forall v \quad |v| < \delta \Rightarrow |h(v)| \le |f(v)||g(v)| \le \varepsilon_1\varepsilon_2 |v|^2 \le \varepsilon_1\varepsilon_2\delta|v| \le \varepsilon|v|$$

so h is vanishing. ∎

LEMMA 16. In $\mathcal{L}(V,V)$, if $x_n \to X$ and $y_n \to Y$ as $n \to \infty$ then $x_n y_n \to XY$ as $n \to \infty$.

Proof. For any $\varepsilon > 0$, set $\varepsilon' = \min\{1, \frac{\varepsilon}{2(|X|+|Y|+1)}\}$; there exists N_1 such that

$$\forall n > N_1 \quad |x_n - X| < \varepsilon',$$

and there exists N_2 such that

$$\forall n > N_2 \quad |y_n - Y| < \varepsilon'.$$

Now, choose $N = \max\{N_1, N_2\}$. Then

$$\begin{aligned}\forall n > N \quad |x_n y_n - XY| &= |x_n(y_n - Y) + (x_n - X)Y| \le |x_n||y_n - Y| + |x_n - X||Y|\\ &\le (|X| + \varepsilon')\varepsilon' + \varepsilon'|Y| \le \varepsilon'(|X| + |Y| + 1) \le \frac{\varepsilon}{2} < \varepsilon,\end{aligned}$$

so $x_n y_n \to XY$ as $n \to \infty$, as required. ∎

LEMMA 17. (Cauchy product) In $\mathcal{L}(V,V)$, if $\sum_{n=0}^{\infty} a_n$ and $\sum_{n=0}^{\infty} b_n$ converge absolutely then $\sum_{n=0}^{\infty} c_n$ also converges absolutely, where $c_n = \sum_{i=0}^{n} a_i b_{n-i}$, and

$$\sum_{n=0}^{\infty} a_n \sum_{n=0}^{\infty} b_n = \sum_{n=0}^{\infty} c_n.$$

Proof. Let $A_0 = \sum_{n=0}^{\infty} |a_n|$, $B_0 = \sum_{n=0}^{\infty} |b_n|$, $A = \sum_{n=0}^{\infty} a_n$, $B = \sum_{n=0}^{\infty} b_n$. Assume $A_0 > 0$ and $B_0 > 0$, since otherwise the result is immediate. First I shall show that $\sum_{n=0}^{\infty} c_n$ converges absolutely. For any n, we have

$$\sum_{k=0}^{n} |c_k| = \sum_{k=0}^{n} \left| \sum_{i=0}^{k} a_i b_{k-i} \right| \le \sum_{k=0}^{n} \sum_{i=0}^{k} |a_i||b_{k-i}| \le \sum_{i=0}^{n} \sum_{j=0}^{n} |a_i||b_j| = \sum_{i=0}^{n} |a_i| \sum_{j=0}^{n} |b_j| \le A_0 B_0,$$

so $\sum_{k=0}^{\infty} |c_k|$ exists, as required.

Next I shall derive the formula for $\sum_{n=0}^{\infty} c_n$. For any n,

$$\left| \sum_{k=0}^{n} c_k - \sum_{i=0}^{n} a_i \sum_{j=0}^{n} b_j \right| = \left| \sum_{(i,j)\in S} a_i b_j \right| \le \sum_{(i,j)\in S} |a_i||b_j|$$

where $S = \{ (i,j) \mid 0 \leq i \leq n, 0 \leq j \leq n, i+j > n \}$. Now, if $i,j \leq \lfloor \frac{n}{2} \rfloor$ then $2i, 2j \leq n$, so $i+j \leq n$, so $(i,j) \notin S$. Thus

$$S \subseteq \big(\{1,\ldots n\} \times \{1,\ldots n\}\big) \setminus \big(\{1,\ldots \lfloor \tfrac{n}{2} \rfloor\} \times \{1,\ldots \lfloor \tfrac{n}{2} \rfloor\}\big)$$

and hence

$$\sum_{(i,j)\in S} |a_i||b_j| \leq \sum_{i,j=0}^{n} |a_i||b_j| - \sum_{i,j=0}^{\lfloor \frac{n}{2} \rfloor} |a_i||b_j| = \sum_{i=0}^{n} |a_i| \sum_{j=0}^{n} |b_j| - \sum_{i=0}^{\lfloor \frac{n}{2} \rfloor} |a_i| \sum_{j=0}^{\lfloor \frac{n}{2} \rfloor} |b_j|.$$

For any $\varepsilon > 0$, set $\varepsilon' = \min\{A_0, B_0, \frac{\varepsilon}{A_0+B_0+1}\} > 0$. We know that there exists N_1 such that

$$\forall n > N_1 \quad A_0 - \varepsilon' < \sum_{i=1}^{n} |a_i| \leq A_0$$

and there exists N_2 such that

$$\forall n > N_2 \quad B_0 - \varepsilon' < \sum_{j=1}^{n} |b_j| \leq B_0.$$

We also have $\left(\sum_{i=1}^{n} a_i\right)\left(\sum_{j=1}^{n} b_j\right) \to AB$ as $n \to \infty$, by lemma 16, so there exists N_3 such that

$$\forall n > N_3 \quad \left| \sum_{i=1}^{n} a_i \sum_{j=1}^{n} b_j - AB \right| < \varepsilon'.$$

Choose $N = \max(2N_1+1, 2N_2+1, N_3)$. Then, for any $n > N$, we have $\lfloor \frac{n}{2} \rfloor > N_1, N_2$, so

$$\begin{aligned}
\left| \sum_{k=0}^{n} c_k - AB \right| &\leq \left| \sum_{k=0}^{n} c_k - \sum_{i=1}^{n} a_i \sum_{j=1}^{n} b_j \right| + \left| \sum_{i=1}^{n} a_i \sum_{j=1}^{n} b_j - AB \right| \\
&\leq \sum_{i=0}^{n} |a_i| \sum_{j=0}^{n} |b_j| - \sum_{i=0}^{\lfloor \frac{n}{2} \rfloor} |a_i| \sum_{j=0}^{\lfloor \frac{n}{2} \rfloor} |b_j| + \varepsilon' \\
&< A_0B_0 - (A_0 - \varepsilon')(B_0 - \varepsilon') + \varepsilon' \\
&= \varepsilon'(A_0 + B_0 - \varepsilon' + 1) < \varepsilon'(A_0 + B_0 + 1) \leq \varepsilon.
\end{aligned}$$

Thus $\sum_{k=0}^{\infty} c_k = AB$, as required. ▮

2.5 Application of the theory to $\mathbf{R}^n$ and matrices

$\mathbf{R}^n$ is a finite-dimensional vector space. If we choose a basis consisting of unit vectors then this induces the norm

$$\left\| \begin{pmatrix} x^1 \\ \vdots \\ x^n \end{pmatrix} \right\| = \sqrt{\sum_{i=1}^{n} (x^i)^2}.$$

Therefore all the above theory for finite-dimensional vector spaces applies to $\mathbf{R}^n$ with this basis and norm.

Let $\mathcal{M}_{nm}$ be the vector space of all $n \times m$ matrices. $\mathcal{M}_{nm}$ is isomorphic to $\mathcal{L}(\mathbf{R}^m, \mathbf{R}^n)$ in the obvious way: a matrix A with components $A^j{}_i$ corresponds to the linear transformation $\sum_{i=1}^m \sum_{j=1}^n A^j{}_i h^i{}_j$. Hence we can transfer across from $\mathcal{L}(\mathbf{R}^m, \mathbf{R}^n)$ to $\mathcal{M}_{nm}$ the basis and the induced norm, giving

$$|A| = \sqrt{\sum_{i=1}^{m} \sum_{j=1}^{n} (A^j{}_i)^2}.$$

Thanks to the isomorphism, lemmas 10–17 continue to hold when the $\mathcal{L}(V, W)$ spaces are replaced by $\mathcal{M}_{nm}$ spaces. In the following sections I shall apply these lemmas both to $\mathcal{L}(V, W)$ spaces and to $\mathcal{M}_{nm}$ spaces.

2.6 Notation

In the following sections I shall sometimes use the notation $o(x)$ to mean a term of the form $R(x)$, for some vanishing function R, and the notation $O(x)$ to mean a term of the form $R(x)$, for some limited function R.

3. Affine Transformations

3.1 Affine transformations and affine matrices

Any two-dimensional affine transformation $G\colon \mathbf{R}^2 \to \mathbf{R}^2$ may be represented by a matrix

$$\overline{G} = \begin{pmatrix} h_{11} & h_{12} & t_1 \\ h_{21} & h_{22} & t_2 \\ 0 & 0 & 1 \end{pmatrix}$$

such that, if we represent a point $(x_1, x_2) \in \mathbf{R}^2$ by a column vector

$$\overline{(x_1, x_2)} = \begin{pmatrix} x_1 \\ x_2 \\ 1 \end{pmatrix},$$

then application of an affine transformation to a point is represented by matrix multiplication: $\overline{G(x)} = \overline{G}\,\overline{x}$. The homogeneous part of the transformation is represented by $h_{11}, h_{12}, h_{21}, h_{22}$, and the translation part is represented by t_1, t_2. I shall refer to such a matrix $\overline{G}$ as an *affine matrix*. I shall write the combination of two affine transformations as $G_1 \cdot G_2$, but shall represent matrix multiplication by juxtaposition; thus, $\overline{G_1 \cdot G_2} = \overline{G_1}\,\overline{G_2}$.

3.2 The Lie group and the Lie algebra

Let $\mathcal{G}$ be the Lie group of all non-singular two-dimensional affine transformations and let $\mathcal{A}$ be the corresponding Lie algebra. Any member A of $\mathcal{A}$ may be represented by a matrix $\overline{A}$ of the form

$$\overline{A} = \begin{pmatrix} h_{11} & h_{12} & t_1 \\ h_{21} & h_{22} & t_2 \\ 0 & 0 & 0 \end{pmatrix}$$

These matrices are called *affine generator matrices*. Let $\mathcal{M}_{33}$ be the vector space of all 3×3 matrices, and let $\mathcal{A}'$ be the subspace consisting of all affine generator matrices. $\mathcal{A}'$ is closed under the commutator operation $[M, N] = MN - NM$ and so can be considered as a Lie algebra; we can use this to define the Lie bracket $[A, B]$ on $\mathcal{A}$,

$$\overline{[A, B]} = [\overline{A}, \overline{B}] = \overline{A}\,\overline{B} - \overline{B}\,\overline{A}.$$

A norm is defined on $\mathcal{M}_{33}$ by

$$\forall M \in \mathcal{M}_{33} \quad |M| = \sqrt{\sum_{a,b=1}^{3} (\mathcal{M}_{ab})^2},$$

and this induces a norm on $\mathcal{A}'$ and a norm on $\mathcal{A}$:

$$\forall A \in \mathcal{A} \quad |A| = |\overline{A}| = \sqrt{\sum_{a=1}^{2} \sum_{b=1}^{3} (\overline{A}_{ab})^2}.$$

Thus $\mathcal{A}$ and $\mathcal{A}'$ are isomorphic as Lie algebras and normed vector spaces.

There is an exponential function $\exp\colon \mathcal{M}_{33} \to \mathcal{M}_{33}$, defined by

$$\forall M \in \mathcal{M}_{33} \quad \exp M = \sum_{n=0}^{\infty} \frac{M^n}{n!}.$$

This series converges absolutely for all M. Note that, for any affine generator matrix M, $\exp M$ is an affine matrix. The inverse function log is defined on a subset of $\mathcal{M}_{33}$ including all affine matrices such that

$$h_{11}h_{22} - h_{12}h_{21} > 0 \quad \text{and} \quad h_{11} + h_{22} > -2\sqrt{h_{11}h_{22} - h_{12}h_{21}}.$$

log maps all such affine matrices to affine generator matrices. For all $M \in \mathcal{M}_{33}$ with $|M| < 1$, $\log(I+M)$ exists and is given by the absolutely convergent power series

$$\log(I+M) = \sum_{n=1}^{\infty} \frac{(-1)^{n+1}M^n}{n},$$

where I is the identity matrix.

There is also an exponential function $\exp\colon \mathcal{A} \to \mathcal{G}$, with an inverse function log defined on a subset of $\mathcal{G}$, related to the exp and log functions on $\mathcal{M}_{33}$ by

$$\overline{\exp A} = \exp \overline{A}, \qquad \overline{\log A} = \log \overline{A}.$$

To calculate exp and log efficiently without using power series, we first consider the problem of exponentiating 2×2 matrices.

3.3 Exponential of a 2×2 matrix

Let X be any 2×2 matrix. Let $X_0 = X - \frac{1}{2}\operatorname{tr}(X)I$, the traceless part of X.

<u>Case 1: $det(X_0) > 0$.</u> Then $\exp X = e^{\operatorname{tr}(X)/2}(\frac{\sin\theta}{\theta} \cdot X_0 + \cos\theta \cdot I)$, where $\theta = \sqrt{det(X_0)}$.

<u>Case 2: $det(X_0) = 0$.</u> Then $\exp X = e^{\operatorname{tr}(X)/2}(X_0 + I)$.

<u>Case 3: $det(X_0) < 0$.</u> Then $\exp X = e^{\operatorname{tr}(X)/2}(\frac{\sinh\theta}{\theta} \cdot X_0 + \cosh\theta \cdot I)$, where $\theta = \sqrt{-det(X_0)}$.

Note that exp is injective on the set of matrices X with $det(X - \frac{1}{2}\operatorname{tr}(X)I) < \pi^2$.

3.4 Logarithm of a 2×2 matrix

Let Y be any 2×2 matrix satisfying $det(Y) > 0$ and $\operatorname{tr}(Y) > -2\sqrt{det(Y)}$. Let $Y_1 = Y/\sqrt{det(Y)}$.

<u>Case 1: $-2 < \operatorname{tr}(Y_1) < 2$.</u> Then $\log Y = \frac{\theta}{\sin\theta}(Y_1 - \frac{1}{2}\operatorname{tr}(Y_1)I) + \frac{1}{2}\log(det(Y))I$, where $\theta = \cos^{-1}(\frac{1}{2}\operatorname{tr}(Y_1)) \in (0,\pi)$.

<u>Case 2: $\operatorname{tr}(Y_1) = 2$.</u> Then $\log Y = Y_1 - \frac{1}{2}\operatorname{tr}(Y_1)I + \frac{1}{2}\ln(det(Y))I$.

<u>Case 3: $2 < \operatorname{tr}(Y_1)$.</u> Then $\log Y = \frac{\theta}{\sinh\theta}(Y_1 - \frac{1}{2}\operatorname{tr}(Y_1)I) + \frac{1}{2}\ln(det(Y))I$, where $\theta = \cosh^{-1}(\frac{1}{2}\operatorname{tr}(Y_1))$.

3.5 Exponential of an affine generator matrix

Let M be any affine generator matrix. Thus M is of the form

$$M = \left(\begin{array}{cc|c} \multicolumn{2}{c|}{X} & x \\ \hline 0 & 0 & 0 \end{array}\right)$$

where X is 2×2 and x is 2×1. Then

$$\exp M = \left(\begin{array}{cc|c} \multicolumn{2}{c|}{\exp X} & y \\ \hline 0 & 0 & 1 \end{array}\right)$$

where $\exp X$ is as given above and y is given as follows.

Case 1: $det(X) \neq 0$. Then $y = (\exp X - I)X^{-1}x$.

Case 2: $det(X) = 0$ and $\mathrm{tr}(X) \neq 0$. Then $y = (I + \frac{e^{\mathrm{tr}(X)} - 1 - \mathrm{tr}(X)}{\mathrm{tr}(X)^2} X)x$.

Case 3: $det(X) = 0$ and $\mathrm{tr}(X) = 0$. Then $y = (I + \frac{1}{2}X)x$.

3.6 Logarithm of an affine matrix

Let M be any affine matrix. Thus M is of the form

$$M = \left(\begin{array}{cc|c} \multicolumn{2}{c|}{Y} & y \\ \hline 0 & 0 & 1 \end{array}\right)$$

where Y is 2×2 and y is 2×1. Then $\log M$ exists iff $\log Y$ exists, in which case

$$\log M = \left(\begin{array}{cc|c} \multicolumn{2}{c|}{X} & x \\ \hline 0 & 0 & 0 \end{array}\right)$$

where $X = \log Y$ and x is given as follows.

Case 1: $det(X) \neq 0$. Then $x = X(Y - I)^{-1}y$.

Case 2: $det(X) = 0$ and $\mathrm{tr}(X) \neq 0$. Then $x = (I + (\frac{1}{e^{\mathrm{tr}(X)} - 1} - \frac{1}{\mathrm{tr}(X)})X)y$.

Case 3: $det(X) = 0$ and $\mathrm{tr}(X) = 0$. Then $x = (I - \frac{1}{2}X)y$.

LEMMA 1. $\log(\exp A) = A$ for all $A \in \mathcal{A}$ with $|A| < \sqrt{2}\,\pi$.

Proof. For any $A \in \mathcal{A}$ with $|A| < \sqrt{2}\,\pi$, we have

$$\overline{A} = \left(\begin{array}{cc|c} \multicolumn{2}{c|}{X} & x \\ \hline 0 & 0 & 0 \end{array}\right), \text{ with } X = \begin{pmatrix} h_{11} & h_{12} \\ h_{21} & h_{22} \end{pmatrix}.$$

and hence

$$det(X - \tfrac{1}{2}\,\mathrm{tr}(X)I) = -\tfrac{1}{4}(h_{11} - h_{22})^2 - h_{12}h_{21} \leq -h_{12}h_{21} \leq \tfrac{1}{2}(h_{12}^2 + h_{21}^2) \leq \tfrac{1}{2}|A|^2 < \pi^2.$$

Hence $\log(\exp X) = X$, by the log construction for 2×2 matrices, and $\log(\exp A) = A$ by the log construction for affine matrices. ∎

3.7 The dual vector space, $\mathcal{F}$

Let $\mathcal{F}$ be the dual vector space to $\mathcal{A}$: that is, $\mathcal{F}$ consists of all linear transformations from $\mathcal{A}$ to $\mathbf{R}$.

Given any basis $\{a_1, \dots a_6\}$ for $\mathcal{A}$, we can define a dual basis $\{f^1, \dots f^6\}$ for $\mathcal{F}$ by

$$f^i(a_j) = \begin{cases} 1 & \text{if } i = j, \\ 0 & \text{if } i \neq j. \end{cases}$$

Any vectors $A \in \mathcal{A}$ and $F \in \mathcal{F}$ can be expressed in terms of the bases as $A = \sum_{i=1}^{6} A^i a_i$ and $F = \sum_{i=1}^{6} F_i f^i$, for unique sequences of real numbers $A^1, \dots A^6$ and $F_1, \dots F_6$. Then we have $F_i = F(a_i)$, $A^i = f^i(A)$, and $F(A) = \sum_{i=1}^{6} F_i A^i$.

For example, one possible basis $\{a_1, \dots a_6\}$ for $\mathcal{A}$ is defined by

$$\overline{a_1} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix} \quad \overline{a_2} = \begin{pmatrix} 0 & 1 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix} \quad \overline{a_3} = \begin{pmatrix} 0 & 0 & 1 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}$$
$$\overline{a_4} = \begin{pmatrix} 0 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix} \quad \overline{a_5} = \begin{pmatrix} 0 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 0 \end{pmatrix} \quad \overline{a_6} = \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 0 \end{pmatrix}.$$

Then any element $A \in \mathcal{A}$, where

$$\overline{A} = \begin{pmatrix} h_{11} & h_{12} & t_1 \\ h_{21} & h_{22} & t_2 \\ 0 & 0 & 0 \end{pmatrix},$$

may be expressed uniquely in terms of this basis as $A = \sum_{i=1}^{6} A^i a_i$, where

$$A^1 = h_{11}, \quad A^2 = h_{12}, \quad A^3 = t_1, \quad A^4 = h_{21}, \quad A^5 = h_{22}, \quad A^6 = t_2.$$

The corresponding basis for $\mathcal{F}$ is $\{f^1, \dots f^6\}$, where

$$f^1(A) = h_{11}, \quad f^2(A) = h_{12}, \quad f^3(A) = t_1,$$
$$f^4(A) = h_{21}, \quad f^5(A) = h_{22}, \quad f^6(A) = t_2,$$

and any $F \in \mathcal{F}$ operates on any $A \in \mathcal{A}$ by the rule

$$F(A) = \sum_{i=1}^{6} F_i A^i = F_1 h_{11} + F_2 h_{12} + F_3 t_1 + F_4 h_{21} + F_5 h_{22} + F_6 t_2.$$

Using this basis, F can be represented by a matrix

$$\overline{F} = \begin{pmatrix} F_1 & F_4 & a \\ F_2 & F_5 & b \\ F_3 & F_6 & c \end{pmatrix}$$

where a, b, c are arbitrary; when I say that $\overline{F}$ represents F, what I mean is that $\overline{F}$ and F are related by the formula

$$\forall A \in \mathcal{A} \quad F(A) = \operatorname{tr}(\overline{F}\,\overline{A}).$$

The norm induced on $\mathcal{A}$ by $\{a_1, \ldots a_6\}$ (see §2.1) is the one we have already defined:

$$|A|^2 = \Big|\sum_{i=1}^{6} A^i a_i\Big|^2 = \sum_{i=1}^{6} (A^i)^2 = \sum_{i=1}^{6} f^i(A)^2.$$

Similarly, a norm is induced on $\mathcal{F}$ by $\{f^1, \ldots f^6\}$:

$$|F|^2 = \Big|\sum_{i=1}^{6} F_i f^i\Big|^2 = \sum_{i=1}^{6} (F_i)^2 = \sum_{i=1}^{6} F(a_i)^2.$$

Given any linear transformation $T: \mathcal{A} \to \mathcal{A}$ the *adjoint* linear transformation $T^\dagger: \mathcal{F} \to \mathcal{F}$ is defined by

$$\forall F \in \mathcal{F}\ \forall A \in \mathcal{A} \quad T^\dagger(F)(A) = F(T(A)).$$

The basis $\{a_1, \ldots a_6\}$ induces a basis on $\mathcal{L}(\mathcal{A}, \mathcal{A})$, the vector space of linear transformations from $\mathcal{A}$ to $\mathcal{A}$, as described in §2.4. This basis induces a norm on $\mathcal{L}(\mathcal{A}, \mathcal{A})$:

$$|T|^2 = \sum_{i,j=1}^{6} (f^j(T(a_i)))^2 = \sum_{i=1}^{6} |T(a_i)|^2.$$

Likewise the basis $\{f^1, \ldots f^6\}$ induces a basis on $\mathcal{L}(\mathcal{F}, \mathcal{F})$, which induces a norm on $\mathcal{L}(\mathcal{F}, \mathcal{F})$:

$$|S|^2 = \sum_{i,j=1}^{6} (S(f^i)(a_j))^2 = \sum_{i=1}^{6} |S(f^i)|^2.$$

LEMMA 2. For any $T \in \mathcal{L}(\mathcal{A}, \mathcal{A})$, $|T^\dagger| = |T|$.

Proof. $|T^\dagger|^2 = \sum_{i,j=1}^{6} (T^\dagger(f^i)(a_j))^2 = \sum_{i,j=1}^{6} (f^i(T(a_j)))^2 = |T|^2$. ∎

3.8 Inner automorphisms of the Lie algebra

Define $Ad: \mathcal{G} \to (\mathcal{A} \to \mathcal{A})$, in terms of matrix representations, by

$$\forall G \in \mathcal{G}\ \forall A \in \mathcal{A} \quad \overline{Ad(G)(A)} = \overline{G}\,\overline{A}\,\overline{G}^{-1}.$$

Note that, for any $G \in \mathcal{G}$, $Ad(G)$ is a linear transformation on $\mathcal{A}$, and that $Ad(G_1) \circ Ad(G_2) = Ad(G_1 G_2)$.

LEMMA 3. For any $G \in \mathcal{G}$ and $A \in \mathcal{A}$, $\exp(Ad(G)(A)) = G \cdot \exp(A) \cdot G^{-1}$.

Proof. In terms of the matrix representations,

$$\begin{aligned}
\overline{\exp(Ad(G)(A))} &= \exp \overline{Ad(G)(A)} = \exp(\overline{G}\,\overline{A}\,\overline{G}^{-1}) \\
&= \sum_{n=0}^{\infty} \frac{1}{n!} (\overline{G}\,\overline{A}\,\overline{G}^{-1})^n = \overline{G} \Big(\sum_{n=0}^{\infty} \frac{1}{n!} \overline{A}^n \Big) \overline{G}^{-1} \\
&= \overline{G} \exp(\overline{A}) \overline{G}^{-1} = \overline{G}\ \overline{\exp(A)}\ \overline{G}^{-1} = \overline{G \cdot \exp(A) \cdot G^{-1}}.
\end{aligned}$$

∎

LEMMA 4. For any $G \in \mathcal{G}$ and $F \in \mathcal{F}$, if $\overline{F}$ is a matrix representing F then the matrix $\overline{G}^{-1}\overline{F}\,\overline{G}$ represents $Ad(G)^{\dagger}(F)$.

Proof. For any $A \in \mathcal{A}$,

$$Ad(G)^{\dagger}(F)(A) = F(Ad(G)(A)) = \operatorname{tr}(\overline{F}\,\overline{Ad(G)(A)}) = \operatorname{tr}(\overline{F}\,\overline{G}\,\overline{A}\,\overline{G}^{-1}) = \operatorname{tr}(\overline{G}^{-1}\overline{F}\,\overline{G}\,\overline{A})$$

which shows that $\overline{G}^{-1}\overline{F}\,\overline{G}$ represents $Ad(G)^{\dagger}(F)$. ∎

Define a linear function $ad: \mathcal{A} \to (\mathcal{A} \to \mathcal{A})$ by

$$\forall A, V \in \mathcal{A} \quad ad(A)(V) = [A, V].$$

Note that, for any A, $ad(A): \mathcal{A} \to \mathcal{A}$ is a linear transformation.

LEMMA 5. For any $A \in \mathcal{A}$, $|ad(A)| \leq \sqrt{5}|A|$.

Proof. Let $\{a_1, \ldots a_6\}$ be the basis that induces the norm on $\mathcal{A}$. The representing matrix $\overline{A}$ of any $A \in \mathcal{A}$ can be written in the form

$$\overline{A} = \begin{pmatrix} h_{11} & h_{12} & t_1 \\ h_{21} & h_{22} & t_2 \\ 0 & 0 & 0 \end{pmatrix}.$$

Then we have

$$\begin{aligned} |ad(A)|^2 &= \sum_{i=1}^{6} |ad(A)(a_i)|^2 = \sum_{i=1}^{6} |[A, a_i]|^2 = \sum_{i=1}^{6} \left|\overline{[A, a_i]}\right|^2 = \sum_{i=1}^{6} \left|[\overline{A}, \overline{a_i}]\right|^2 \\ &= 2(h_{11} - h_{22})^2 + h_{11}^2 + h_{22}^2 + 5h_{12}^2 + 5h_{21}^2 + 2t_1^2 + 2t_2^2 \\ &\leq 5h_{11}^2 + 5h_{22}^2 + 5h_{12}^2 + 5h_{21}^2 + 2t_1^2 + 2t_2^2 \leq 5|\overline{A}|^2 = 5|A|^2 \end{aligned}$$

so $|ad(A)| \leq \sqrt{5}|A|$. ∎

LEMMA 6. For any $A \in \mathcal{A}$ and $F \in \mathcal{F}$, if $\overline{F}$ is a matrix representing F then the matrix $[\overline{F}, \overline{A}]$ represents $ad(A)^{\dagger}(F)$.

Proof. For any $V \in \mathcal{A}$,

$$ad(A)^{\dagger}(F)(V) = F(ad(A)(V)) = \operatorname{tr}(\overline{F}\,\overline{ad(A)(V)}) = \operatorname{tr}(\overline{F}\,\overline{[A, F]}) = \operatorname{tr}(\overline{F}[\overline{A}, \overline{V}]) = \operatorname{tr}([\overline{F}, \overline{A}]\overline{V})$$

which shows that $[\overline{F}, \overline{A}]$ represents $ad(A)^{\dagger}(F)$. ∎

3.9 Approximation lemmas for exp and log

LEMMA 7. For all $A \in \mathcal{A}'$, $\exp A = I + A + R_E(A)$, where $R_E: \mathcal{A}' \to \mathcal{A}'$ is vanishing.

Proof. Define $R: \mathcal{M}_{33} \to \mathcal{M}_{33}$ by $\forall A \in \mathcal{M}_{33}\ R(A) = \exp A - I - A$. By the power series expansion for exp, we have

$$\forall A \in \mathcal{M}_{33} \quad R(A) = \sum_{n=2}^{\infty} \frac{A^n}{n!}.$$

The corresponding real power series $\sum_{n=2}^{\infty} \frac{t^n}{n!}$ is absolutely convergent for all $t \in \mathbf{R}$, so, by lemma 2.14 (transferred to $\mathcal{M}_{33}$, as explained in §2.5), R is vanishing. Note that $R(\mathcal{A}') \subseteq \mathcal{A}'$, so we can define R_E as the restriction of R to $\mathcal{A}'$. ∎

LEMMA 8. For all $A \in \mathcal{A}'$ with $|A| < 1$, $\log(I + A) = A + R_L(A)$, where $R_L: \mathcal{A}' \to \mathcal{A}'$ is vanishing.

Proof. Let N be the set of all $A \in \mathcal{M}_{33}$ such that $|A| < 1$. Define $R: \mathcal{M}_{33} \to \mathcal{M}_{33}$ such that $\forall A \in N \ R(A) = \log(I + A) - A$ (the values of R outside N are arbitrary). By the power series expansion for log, we have

$$\forall A \in N \quad R(A) = \sum_{n=2}^{\infty} \frac{(-1)^{n+1} A^n}{n}.$$

The corresponding real power series $\sum_{n=2}^{\infty} \frac{(-1)^{n+1} t^n}{n}$ is absolutely convergent for all $t \in (-1, 1)$, so, by lemma 2.14 (transferred to $\mathcal{M}_{33}$), R is vanishing. Note that $R(\mathcal{A}') \subseteq \mathcal{A}'$, so we can define R_L as the restriction of R to $\mathcal{A}'$. ∎

LEMMA 9. If S is any finite-dimensional vector space, $F: S \to \mathcal{A}'$ is linear, and $R_1: S \to \mathcal{A}'$ is vanishing, then there is a vanishing function $R_2: S \to \mathcal{A}'$ such that, for all $V \in S$ in a neighbourhood of 0,

$$I + F(V) + R_1(V) = \exp(F(V) + R_2(V)).$$

Proof. F is linear and hence limited, R_1 is vanishing and hence limited, so $F + R_1$ is limited and hence continuous at 0, with $(F + R_1)(0) = 0$ (lemmas 2.3, 2.4, 2.5, 2.7). Therefore there is a neighbourhood N of 0 for which

$$\forall V \in N \quad |F(V) + R_1(V)| < 1.$$

For all $V \in N$, we have

$$\log(I + F(V) + R_1(V)) = F(V) + R_1(V) + R_L(F(V) + R_1(V))$$

by lemma 8. Define $R_2 = R_1 + R_L \circ (F + R_1)$, giving

$$\forall V \in N \quad I + F(V) + R_1(V) = \exp(F(V) + R_2(V)).$$

Now, R_L is vanishing and $F + R_1$ is limited, so $R_L \circ (F + R_1)$ is vanishing, so R_2 is vanishing (lemmas 2.6, 2.8), as required. ∎

LEMMA 10. For any $A, B \in \mathcal{A}$, there is a vanishing function $R: \mathbf{R} \to \mathcal{A}$, such that, for all real ε in a neighbourhood of 0,

$$\exp(\varepsilon A) \cdot \exp(\varepsilon B) = \exp(\varepsilon(A + B) + R(\varepsilon)).$$

Proof. By lemma 7,

$$\forall \varepsilon \quad \exp(\varepsilon \overline{A}) \exp(\varepsilon \overline{B}) = (I + \varepsilon \overline{A} + R_E(\varepsilon \overline{A}))(I + \varepsilon \overline{B} + R_E(\varepsilon \overline{B})) = I + \varepsilon(\overline{A} + \overline{B}) + R_1(\varepsilon)$$

where $R_1: \mathbf{R} \to \mathcal{A}'$ is defined by

$$\forall \varepsilon \quad R_1(\varepsilon) = R_E(\varepsilon \overline{A}) + R_E(\varepsilon \overline{B}) + \varepsilon^2 \overline{A}\,\overline{B} + \varepsilon \overline{A}\, R_E(\varepsilon \overline{B}) + R_E(\varepsilon \overline{A})\, \varepsilon \overline{B} + R_E(\varepsilon \overline{A}) R_E(\varepsilon \overline{B}).$$

If we define linear functions $L_1, L_2 : \mathbf{R} \to \mathcal{A}'$ by

$$\forall \varepsilon \quad L_1(\varepsilon) = \varepsilon \overline{A}, \quad L_2(\varepsilon) = \varepsilon \overline{B},$$

then L_1, L_2 are limited, so $R_E \circ L_1, R_E \circ L_2$ are vanishing, so R_1 is vanishing (lemmas 2.3, 2.4, 2.6, 2.8, 2.15).

Then, by lemma 9, there is a vanishing function $R_2 : \mathbf{R} \to \mathcal{A}'$ such that, for all ε in a neighbourhood of 0,

$$I + \varepsilon(\overline{A} + \overline{B}) + R_1(\varepsilon) = \exp(\varepsilon(\overline{A} + \overline{B}) + R_2(\varepsilon))$$

giving

$$\exp(\varepsilon \overline{A}) \exp(\varepsilon \overline{B}) = \exp(\varepsilon(\overline{A} + \overline{B}) + R_2(\varepsilon)).$$

Using the isomorphism between $\mathcal{A}'$ and $\mathcal{A}$ we can obtain from R_2 a vanishing function $R : \mathbf{R} \to \mathcal{A}$ such that, for all ε in a neighbourhood of 0,

$$\exp(\varepsilon A) \cdot \exp(\varepsilon B) = \exp(\varepsilon(A + B) + R(\varepsilon))$$

as required. ∎

3.10 Derivatives of functions to and from $\mathcal{G}$

Let S be any finite-dimensional vector space.

The *derivative* of a partial function $f : \mathcal{G} \to S$ is the unique maximal partial function $f_* : \mathcal{G} \to (\mathcal{A} \to S)$ such that, for every $G \in \mathcal{G}$ for which $f_*(G)$ is defined, $f_*(G)$ is a linear transformation and there is a vanishing function $R : \mathcal{A} \to S$ such that, for all $V \in \mathcal{A}$ in a neighbourhood of 0,

$$f(G \cdot \exp V) = f(G) + f_*(G)(V) + R(V).$$

f is *differentiable at* G iff $f_*(G)$ is defined.

The *derivative* of a partial function $f : S \to \mathcal{G}$ is the unique maximal partial function $f_* : S \to (S \to \mathcal{A})$ such that, for every $x \in S$ for which $f_*(x)$ is defined, $f_*(x)$ is a linear transformation and there is a vanishing function $R : S \to \mathcal{A}$ such that, for all $v \in S$ in a neighbourhood of 0,

$$f(x + v) = f(x) \cdot \exp(f_*(x)(v) + R(v)).$$

f is *differentiable at* x iff $f_*(x)$ exists.

LEMMA 11. If $G \in \mathcal{G}$, $f : \mathcal{G} \to S$ is differentiable at G, and $R_1 : \mathbf{R} \to \mathcal{A}$ is vanishing, then there is a vanishing function $R_2 : \mathbf{R} \to S$ such that, for all $V \in \mathcal{A}$ and for all real ε in a neighbourhood of 0,

$$f(G \cdot \exp(\varepsilon V + R_1(\varepsilon))) = f(G) + \varepsilon f_*(G)(V) + R_2(\varepsilon).$$

Proof. Since f is differentiable at G, there is a vanishing function $R : \mathcal{A} \to S$ and a neighbourhood $N \subseteq \mathcal{A}$ of 0 such that

$$\forall A \in N \quad f(G \cdot \exp A) = f(G) + f_*(G)(A) + R(A).$$

Given any $V \in \mathcal{A}$, define the linear function $L: \mathbf{R} \to \mathcal{A}$ by $\forall \varepsilon \in \mathbf{R}\ L(\varepsilon) = \varepsilon V$. Now, L and R_1 are limited, so $L + R_1$ is limited and hence continuous at 0, with $(L + R_1)(0) = 0$ (lemmas 2.3, 2.4, 2.5, 2.7). Hence there is a neighbourhood $N' \subseteq \mathbf{R}$ of 0 such that, for all $\varepsilon \in N'$, we have $\varepsilon V + R_1(\varepsilon) = (L + R_1)(\varepsilon) \in N$ and hence

$$f(G \cdot \exp(\varepsilon V + R_1(\varepsilon))) = f(G) + f_*(G)(\varepsilon V + R_1(\varepsilon)) + R(\varepsilon V + R_1(\varepsilon)) = f(G) + \varepsilon f_*(G)(V) + R_2(\varepsilon)$$

where $R_2 = f_*(G) \circ R_1 + R \circ (L + R_1)$. Now, $f_*(G)$ is linear and hence limited, so $f_*(G) \circ R_1$ is vanishing. Also, $R \circ (L + R_1)$ is vanishing. Thus R_2 is vanishing, as required (lemmas 2.4, 2.6, 2.8, 2.9). ∎

3.11 Calculation of the derivative of exp in terms of matrices

For this section only, we shall need a function $ad: \mathcal{A}' \to (\mathcal{A}' \to \mathcal{A}')$ defined by $\forall A, V \in \mathcal{A}'\ ad(A)(V) = [A, V]$, analogous to the function $ad: \mathcal{A} \to (\mathcal{A} \to \mathcal{A})$ already defined.

LEMMA 12. For any natural number n and $A, V \in \mathcal{A}'$,

$$VA^n = \sum_{i=0}^{n} (-1)^i \binom{n}{i} A^{n-i}\, ad(A)^i(V).$$

Proof. By induction on n. ∎

THEOREM 13. For any $A \in \mathcal{A}$,

$$\exp_*(A) = \sum_{n=0}^{\infty} \frac{(-1)^n ad(A)^n}{(n+1)!}$$

and thus, for any $A, V \in \mathcal{A}$,

$$\exp_*(A)(V) = \sum_{n=0}^{\infty} \frac{(-1)^n ad(A)^n(V)}{(n+1)!} = V - \tfrac{1}{2}[A, V] + \tfrac{1}{6}[A, [A, V]] - \tfrac{1}{24}[A, [A, [A, V]]] + \cdots.$$

Both series converge absolutely for all A and V.

Proof. I shall prove the theorem for $A, V \in \mathcal{A}'$ and then transfer the result to $A, V \in \mathcal{A}$, using the isomorphism between $\mathcal{A}'$ and $\mathcal{A}$.

For any $A, V \in \mathcal{A}'$,

$$\exp(A + V) = \sum_{n=0}^{\infty} \frac{1}{n!}(A + V)^n = \sum_{n=0}^{\infty} (X_n(A) + Y_n(A, V) + Z_n(A, V))$$

where $X_n(A) = \frac{1}{n!}A^n$, $Y_n(A, V) = \frac{1}{n!}\sum_{r=1}^{n} A^{n-r} V A^{r-1}$, and $Z_n(A, V)$ is the sum of the remaining terms in the expansion of $\frac{1}{n!}(A + V)^n$. Note that $Z_n(A, V)$ consists of a sum of products of A and V, with positive coefficients. For every n we have a similar expansion of $\frac{1}{n!}(|A| + |V|)^n$:

$$\frac{1}{n!}(|A| + |V|)^n = X_n(|A|) + Y_n(|A|, |V|) + Z_n(|A|, |V|)$$

with

$$
\begin{aligned}
|X_n(A)| &\le X_n(|A|) = \frac{1}{n!}|A|^n,\\
|Y_n(A,V)| &\le Y_n(|A|,|V|) \le \frac{1}{n!}(|A|+|V|)^n,\\
|Z_n(A,V)| &\le Z_n(|A|,|V|) \le \frac{1}{n!}(|A|+|V|)^n.
\end{aligned}
$$

Hence $\sum_{n=0}^{\infty} X_n(A)$, $\sum_{n=0}^{\infty} Y_n(A,V)$, and $\sum_{n=0}^{\infty} Z_n(A,V)$ all converge absolutely, by comparison with $\sum_{n=0}^{\infty} \frac{1}{n!}(|A|+|V|)^n = e^{|A|+|V|}$. Hence

$$
\exp(A+V) = \sum_{n=0}^{\infty} X_n(A) + \sum_{n=0}^{\infty} Y_n(A,V) + \sum_{n=0}^{\infty} Z_n(A,V).
$$

We already know the sum of the first series: $\sum_{n=0}^{\infty} X_n(A) = \exp A$. As for the third series, for fixed A, define $R_A\colon \mathcal{A}' \to \mathcal{A}'$ by

$$
\forall V \in \mathcal{A}' \quad R_A(V) = \sum_{n=0}^{\infty} Z_n(A,V).
$$

I shall show that R_A is vanishing. As pointed out above, we have

$$
\begin{aligned}
|Z_n(A,V)| \le Z_n(|A|,|V|) &= \frac{1}{n!}(|A|+|V|)^n - X_n(|A|) - Y_n(|A|,|V|)\\
&= \frac{1}{n!}((|A|+|V|)^n - |A|^n - n|A|^{n-1}|V|)
\end{aligned}
$$

so

$$
\begin{aligned}
|R_A(V)| &\le \sum_{n=0}^{\infty} |Z_n(A,V)| \le \sum_{n=0}^{\infty} \frac{1}{n!}((|A|+|V|)^n - |A|^n - n|A|^{n-1}|V|)\\
&= e^{|A|+|V|} - e^{|A|} - e^{|A|}|V| = e^{|A|} f(|V|)
\end{aligned}
$$

where $f\colon \mathbf{R} \to \mathbf{R}$ is defined by $\forall t \in \mathbf{R}\ f(t) = \sum_{n=2}^{\infty} \frac{t^n}{n!}$. This real power series is absolutely convergent for all t, so, by lemma 2.14 (transferred to $\mathcal{M}_{11}$, i.e., $\mathbf{R}$), f is vanishing. Hence R_A is also vanishing.

Returning to the second series, $\sum_{n=0}^{\infty} Y_n(A,V)$, we have $Y_0(A,V) = 0$ and, for any $n > 0$, by lemma 12,

$$
\begin{aligned}
Y_n(A,V) &= \frac{1}{n!}\sum_{r=1}^{n} A^{n-r} V A^{r-1}\\
&= \frac{1}{n!}\sum_{r=1}^{n}\sum_{i=0}^{r-1} (-1)^i \binom{r-1}{i} A^{n-i-1}\, ad(A)^i(V)\\
&= \frac{1}{n!}\sum_{i=0}^{n-1}\sum_{r=i+1}^{n} (-1)^i \binom{r-1}{i} A^{n-i-1}\, ad(A)^i(V)\\
&= \frac{1}{n!}\sum_{i=0}^{n-1} (-1)^i \binom{n}{i+1} A^{n-i-1}\, ad(A)^i(V)\\
&= \sum_{i=0}^{n-1} \frac{A^{n-i-1}}{(n-i-1)!}\, \frac{(-1)^i ad(A)^i(V)}{(i+1)!}.
\end{aligned}
$$

I wish to apply the Cauchy product formula (lemma 2.17, transferred to $\mathcal{M}_{33}$) to this expression to evaluate $\sum_{n=1}^{\infty} Y_n(A,V)$. This is justified since the series $\sum_{j=0}^{\infty} A^j/j!$ and $\sum_{i=0}^{\infty}(-1)^i ad(A)^i(V)/(i+1)!$ are absolutely convergent: in the case of the latter series we have

$$\left|\frac{(-1)^i ad(A)^i(V)}{(i+1)!}\right| \le \frac{|\sqrt{5}A|^i}{(i+1)!}|V|$$

by lemma 5. The Cauchy product formula gives

$$\begin{aligned}\sum_{n=1}^{\infty} Y_n(A,V) &= \sum_{n=1}^{\infty}\sum_{i=0}^{n-1} \frac{A^{n-i-1}}{(n-i-1)!}\,\frac{(-1)^i ad(A)^i(V)}{(i+1)!} \\ &= \sum_{m=0}^{\infty}\sum_{i=0}^{m} \frac{A^{m-i}}{(m-i)!}\,\frac{(-1)^i ad(A)^i(V)}{(i+1)!} \\ &= \sum_{j=0}^{\infty}\frac{A^j}{j!}\sum_{i=0}^{\infty}\frac{(-1)^i ad(A)^i(V)}{(i+1)!} \\ &= \exp(A)\sum_{i=0}^{\infty}\frac{(-1)^i ad(A)^i(V)}{(i+1)!}.\end{aligned}$$

Putting all three pieces together gives

$$\begin{aligned}\exp(A+V) &= \exp(A) + \exp(A)\sum_{i=0}^{\infty}\frac{(-1)^i ad(A)^i(V)}{(i+1)!} + R_A(V) \\ &= \exp(A)\left(I + \sum_{i=0}^{\infty}\frac{(-1)^i ad(A)^i(V)}{(i+1)!} + E_A(R_A(V))\right)\end{aligned}$$

where $E_A\colon \mathcal{A}' \to \mathcal{A}'$ is the linear transformation defined by $\forall X \in \mathcal{A}'\ E_A(X) = \exp(-A)X$. By lemma 2.4, E_A is limited, so, by lemma 2.9, $E_A \circ R_A$ is vanishing. Hence, by lemma 9, there is a vanishing function $R\colon \mathcal{A}' \to \mathcal{A}'$ such that, for all V in a neighbourhood of 0,

$$\exp(A+V) = \exp(A)\exp\left(\sum_{i=0}^{\infty}\frac{(-1)^i ad(A)^i(V)}{(i+1)!} + R(V)\right).$$

This equation has been proved for $A, V \in \mathcal{A}'$. Since $\mathcal{A}'$ is isomorphic to $\mathcal{A}$ the equation also holds for $A, V \in \mathcal{A}$. This gives $\exp_*(A)(V) = \sum_{i=0}^{\infty}\frac{(-1)^i ad(A)^i(V)}{(i+1)!}$ and hence $\exp_*(A) = \sum_{i=0}^{\infty}\frac{(-1)^i ad(A)^i}{(i+1)!}$ by lemma 2.12, as required. As indicated above, these series are absolutely convergent by comparison with $\sum_{i=0}^{\infty}\frac{|\sqrt{5}A|^i}{(i+1)!}|V|$ and $\sum_{i=0}^{\infty}\frac{|\sqrt{5}A|^i}{(i+1)!}$. ∎

3.12 Calculation of $\log_*$

Define a real sequence (a_n) by

$$a_n = \sum_{m=0}^{n}\ \sum_{\substack{r_1,\ldots r_m \ge 1 \\ r_1+\cdots+r_m=n}} \frac{(-1)^{n-m}}{(r_1+1)!\cdots(r_m+1)!}.$$

LEMMA 14. The real power series $\sum_{n=0}^{\infty} a_n x^n$ converges absolutely if $|x| < 1$.

Proof. We have

$$|a_n| \leq \sum_{m=0}^{n} \sum_{\substack{r_1,\dots r_m \geq 1 \\ r_1+\cdots+r_m=n}} \frac{1}{(r_1+1)!\cdots(r_m+1)!} \leq \sum_{m=0}^{n}\left(\sum_{r=1}^{n}\frac{1}{(r+1)!}\right)^m$$
$$< \sum_{m=0}^{n}(e-2)^m < \frac{1}{1-(e-2)} = \frac{1}{3-e}$$

so the series $\sum_{n=0}^{\infty}|a_n x^n|$ converges by comparison with $\frac{1}{3-e}\sum_{n=0}^{\infty}|x|^n$, if $|x|<1$. ∎

In fact, the convergence of the power series is much better than this: the first few terms of (a_n) are $(1, \frac{1}{2}, \frac{1}{12}, 0, -\frac{1}{720}, 0, \frac{1}{30240}, 0, -\frac{1}{1209600}, \dots)$. The next lemma shows that $\sum_{n=0}^{\infty} a_n x^n$ is the inverse of $\sum_{n=0}^{\infty}\frac{(-1)^n}{(n+1)!}x^n$, in the sense of the Cauchy product of series.

LEMMA 15. For any $p \geq 0$,

$$\sum_{n=0}^{p}\frac{(-1)^n}{(n+1)!}a_{p-n} = \begin{cases} 1 & \text{if } p=0 \\ 0 & \text{if } p>0 \end{cases}$$

Proof. Define $S_n = \frac{(-1)^{n+1}}{(n+1)!}$. Then the definition of a_n may be rewritten as

$$a_n = \sum_{m=0}^{n} b_m^n, \quad \text{where } b_m^n = \sum_{\substack{r_1,\dots r_m \geq 1 \\ r_1+\cdots+r_m=n}} S_{r_1}S_{r_2}\cdots S_{r_m}.$$

Note that $S_0=-1$, $b_0^0=1$, $b_0^p=0$ if $p>0$, and $b_m^n=0$ if $m>n$. Then

$$\begin{aligned}\sum_{n=0}^{p}\frac{(-1)^n}{(n+1)!}a_{p-n} &= \sum_{n=0}^{p}(-S_n)\sum_{m=0}^{p-n}b_m^{p-n} = -\sum_{n=0}^{p}\sum_{m=0}^{p}S_n b_m^{p-n} \\ &= -\sum_{m=0}^{p}S_0 b_m^p - \sum_{n=1}^{p}\sum_{m=0}^{p}S_n b_m^{p-n} = \sum_{m=0}^{p}b_m^p - \sum_{m=0}^{p}\sum_{n=1}^{p}S_n b_m^{p-n} \\ &= \sum_{m=0}^{p}b_m^p - \sum_{m=0}^{p}b_{m+1}^p = b_0^p - b_{p+1}^p = \begin{cases} 1 & \text{if } p=0 \\ 0 & \text{if } p>0 \end{cases}\end{aligned}$$

∎

THEOREM 16. For any $A \in \mathcal{A}$ with $|A| < \frac{1}{\sqrt{5}}$,

$$\log_*(\exp A) = \sum_{n=0}^{\infty} a_n\, ad(A)^n;$$

thus, for any $V \in \mathcal{A}$,

$$\begin{aligned}\log_*(\exp A)(V) &= \sum_{n=0}^{\infty} a_n\, ad(A)^n(V) \\ &= V + \tfrac{1}{2}[A,V] + \tfrac{1}{12}[A,[A,V]] - \tfrac{1}{720}[A,[A,[A,[A,V]]]] \\ &\quad + \tfrac{1}{30240}[A,[A,[A,[A,[A,[A,V]]]]]] \\ &\quad - \tfrac{1}{1209600}[A,[A,[A,[A,[A,[A,[A,[A,V]]]]]]]] + \cdots.\end{aligned}$$

Both series converge absolutely for $|A| < \frac{1}{\sqrt{5}}$.

Proof. For $|A| < \frac{1}{\sqrt{5}}$ we have $|ad(A)| < 1$, by lemma 5, so $\sum_{n=0}^{\infty} a_n \, ad(A)^n$ and $\sum_{n=0}^{\infty} a_n \, ad(A)^n(V)$ converge absolutely by lemma 14 and lemma 2.13. I intend to apply the inverse function theorem, so I must show that exp is continuously differentiable and $\exp_*(A)$ is invertible.

Theorem 13 shows that exp is differentiable, with $\exp_* = F \circ ad$, where $F\colon \mathcal{L}(\mathcal{A}, \mathcal{A}) \to \mathcal{L}(\mathcal{A}, \mathcal{A})$ is defined by

$$\forall T \in \mathcal{L}(\mathcal{A}, \mathcal{A}) \quad F(T) = \sum_{n=0}^{\infty} \frac{(-1)^n}{(n+1)!} T^n.$$

The corresponding real power series $\sum_{n=0}^{\infty} \frac{(-1)^n}{(n+1)!} x^n$ is absolutely convergent for all $x \in \mathbf{R}$, so by lemma 2.13 the power series for F is absolutely convergent for all T and F is continuous. The linear function ad is also continuous, by lemma 2.4, so $\exp_* = F \circ ad$ is continuous.

Since the series $\sum_{n=0}^{\infty} \frac{(-1)^n}{(n+1)!} ad(A)^n$ and $\sum_{n=0}^{\infty} a_n \, ad(A)^n$ are absolutely convergent (for $|A| < \frac{1}{\sqrt{5}}$), we can compose them, using the Cauchy product formula (lemma 2.17) and lemma 15, to give

$$\begin{aligned}
\sum_{n=0}^{\infty} \frac{(-1)^n}{(n+1)!} ad(A)^n \circ \sum_{n=0}^{\infty} a_n \, ad(A)^n &= \sum_{p=0}^{\infty} \left(\sum_{n=0}^{p} \frac{(-1)^n}{(n+1)!} a_{p-n} \right) ad(A)^p \\
&= 1 \cdot ad(A)^0 + 0 \cdot ad(A)^1 + 0 \cdot ad(A)^2 + \cdots = I
\end{aligned}$$

where I is the identity function. This shows that $\exp_*(A)$ is invertible and that $\sum_{n=0}^{\infty} a_n \, ad(A)^n$ is its inverse.

Hence, by the inverse function theorem, exp maps an open neighbourhood of A bijectively onto an open neighbourhood of $\exp A$, and its inverse log is continuously differentiable. It follows that $\log_*(\exp A)$ is the inverse of $\exp_*(A)$, which is $\sum_{n=0}^{\infty} a_n \, ad(A)^n$.

Applying lemma 2.12 gives $\log_*(\exp A)(V) = \sum_{n=0}^{\infty} a_n \, ad(A)^n(V)$. ∎

3.13 The dual mapping, $\log^*$

We can define a function $\log^*\colon \mathcal{G} \to (\mathcal{F} \to \mathcal{F})$ dual to $\log_*\colon \mathcal{G} \to (\mathcal{A} \to \mathcal{A})$ by

$$\forall G \in \mathcal{G} \quad \log^*(G) = (\log_*(G))^\dagger$$

or, more explicitly,

$$\forall G \in \mathcal{G} \ \forall F \in \mathcal{F} \ \forall V \in \mathcal{A} \quad \log^*(G)(F)(V) = F(\log_*(G)(V)).$$

THEOREM 17. For any $A \in \mathcal{A}$ with $|A| < \frac{1}{\sqrt{5}}$,

$$\log^*(\exp A) = \sum_{n=0}^{\infty} a_n (ad(A)^\dagger)^n$$

where the real sequence (a_n) is as above; thus, for any $F \in \mathcal{F}$,

$$\log^*(\exp A)(F) = \sum_{n=0}^{\infty} a_n (ad(A)^\dagger)^n (F).$$

Both series converge absolutely for $|A| < \frac{1}{\sqrt{5}}$.

Proof. For any $V \in \mathcal{A}$ we have

$$\log_*(\exp A)(V) = \sum_{n=0}^{\infty} a_n \, ad(A)^n (V)$$

by theorem 16. So, for any $F \in \mathcal{F}$,

$$F(\log_*(\exp A)(V)) = \sum_{n=0}^{\infty} a_n F(ad(A)^n (V))$$

since F is continuous by lemma 2.4. Hence

$$\log^*(\exp A)(F)(V) = F(\log_*(\exp A)(V)) = \sum_{n=0}^{\infty} a_n F(ad(A)^n (V)) = \sum_{n=0}^{\infty} a_n (ad(A)^\dagger)^n (F)(V).$$

Applying lemma 2.12 gives $\log^*(\exp A)(F) = \sum_{n=0}^{\infty} a_n (ad(A)^\dagger)^n (F)$ and $\log^*(\exp A) = \sum_{n=0}^{\infty} a_n (ad(A)^\dagger)^n$, as required. These series converge absolutely for $|A| < \frac{1}{\sqrt{5}}$ since $|ad(A)^\dagger| = |ad(A)| < 1$, by lemma 2 and lemma 5. ∎

4. Matching a Template to an Image

4.1 Images and templates

An *image* is a function $I: \mathbf{R}^2 \to [0, \infty)$. For any point $p \in \mathbf{R}^2$, $I(p)$ is the image intensity at the pixel p. The domain of I is called the *image plane*.

As mentioned in §1, each symbol type has a *template*, depicting the appearance of any token of that symbol type at a standard position, size and orientation, in the absence of noise. Formally, a template is a differentiable function $T: \mathbf{R}^2 \to [0, \infty)$ such that the set $\{ x \in \mathbf{R}^2 \mid T(x) > 0 \}$ is bounded. The domain of T is called the *template plane*.

Suppose that we believe that a token of this symbol type occurs in the image at a certain position, orientation and size (with a certain degree of stretching and shearing). We can describe this by an affine transformation G from the template plane into the image plane, which is called the *embedding* of the symbol token in the image. If the symbol token really is present where we think it is, there will be a good match between the functions T and $I \circ G$, at least in the region where T is non-zero.

The aim of this section is to measure the goodness of match by a correlation function between T and $I \circ G$, and to calculate its derivative so that we can adjust G incrementally to maximise the correlation.

Recall from §3.1 that an affine transformation G can be represented by a 3×3 matrix $\overline{G}$ and a point $u \in \mathbf{R}^2$ can be represented by a 3×1 column vector $\overline{u}$, such that $\overline{G(u)} = \overline{G}\,\overline{u}$. For any template T we can define a modified function $\overline{T}$ on 3×1 column vectors, such that $\forall u \in \mathbf{R}^2 \ \overline{T}(\overline{u}) = T(u)$.

4.2 The correlation function

Given T, I and G, define the correlation $\rho_{I,T}(G)$ between T and $I \circ G$ by

$$\rho_{I,T}(G) = |det(\overline{G})| \int T(u)\,(I(G(u)) - I_0)\,\mathrm{d}^2 u = \int T(G^{-1}(x))\,(I(x) - I_0)\,\mathrm{d}^2 x$$

where I_0 is a positive real constant associated with T. The integrals are over the whole of $\mathbf{R}^2$, or equivalently over a large enough region to include in its interior all the points where T is non-zero.

LEMMA 1. For any template T, any image I, and any affine transformations G, G',

$$\rho_{I,T\circ G'}(G \cdot G') = \rho_{I,T}(G).$$

Proof.

$$\begin{aligned}
\rho_{I,T\circ G'}(G \cdot G') &= \int (T \circ G')((G \cdot G')^{-1}(x))\,(I(x) - I_0)\,\mathrm{d}^2 x \\
&= \int T(G'(G'^{-1}(G^{-1}(x))))\,(I(x) - I_0)\,\mathrm{d}^2 x \\
&= \int T(G^{-1}(x))\,(I(x) - I_0)\,\mathrm{d}^2 x = \rho_{I,T}(G).
\end{aligned}$$

∎

4.3 Derivative of the correlation function

Let G be an affine transformation, and consider the effect on $\rho_{I,T}(G)$ of making a small change in G, i.e., replacing G by $G \cdot \exp A$, for some $A \in \mathcal{A}$. We have

$$\begin{aligned}
\rho_{I,T}(G \cdot \exp A) &= \int T((G \cdot \exp A)^{-1}(x))\,(I(x) - I_0)\,\mathrm{d}^2 x \\
&= \int \overline{T}((\overline{G} \exp \overline{A})^{-1} \overline{x})\,(I(x) - I_0)\,\mathrm{d}^2 x \\
&= \int \overline{T}(\exp(-\overline{A})\overline{G}^{-1} \overline{x})\,(I(x) - I_0)\,\mathrm{d}^2 x \\
&= \int \overline{T}((I - \overline{A} + o(A))\overline{G}^{-1} \overline{x})\,(I(x) - I_0)\,\mathrm{d}^2 x \\
&= \rho_{I,T}(G) \;-\; \int (\overline{\nabla} T(G^{-1}(x)))\overline{A}\,\overline{G}^{-1} \overline{x}\,(I(x) - I_0)\,\mathrm{d}^2 x \;+\; o(A)
\end{aligned}$$

where $\overline{\nabla} T$ is defined by

$$\overline{\nabla} T(u) = \begin{pmatrix} \frac{\partial T(u)}{\partial u_1} & \frac{\partial T(u)}{\partial u_2} & 0 \end{pmatrix}.$$

Hence the derivative $\rho_{I,T\,*}\colon \mathcal{G} \to (\mathcal{A} \to \mathbf{R})$ is given by

$$\begin{aligned}
\forall G \in \mathcal{G}\ \forall A \in \mathcal{A} \quad \rho_{I,T\,*}(G)(A) &= -\int (\overline{\nabla} T(G^{-1}(x)))\overline{A}\,\overline{G}^{-1} \overline{x}\,(I(x) - I_0)\,\mathrm{d}^2 x \\
&= -|det(\overline{G})| \int (\overline{\nabla} T(u))\overline{A}\,\overline{u}\,(I(G(u)) - I_0)\,\mathrm{d}^2 u.
\end{aligned}$$

$\rho_{I,T\,*}(G)$ is a linear transformation from $\mathcal{A}$ to $\mathbf{R}$, and hence is an element of $\mathcal{F}$; it is called the force exerted by I and T on G.

4.4 Matrix representation of the force

LEMMA 2. Using the matrix representation of $\mathcal{F}$ (see §3.7), the force $\rho_{I,T\,*}(G)$ is represented by the matrix

$$\begin{aligned}
\overline{\rho_{I,T\,*}(G)} &= -|det(\overline{G})| \int \begin{pmatrix} u_1 \\ u_2 \\ 1 \end{pmatrix} \begin{pmatrix} \frac{\partial T(u)}{\partial u_1} & \frac{\partial T(u)}{\partial u_2} & 0 \end{pmatrix} (I(G(u)) - I_0)\,\mathrm{d}^2 u \\
&= -|det(\overline{G})| \int \overline{u}\,(\overline{\nabla} T(u))\,(I(G(u)) - I_0)\,\mathrm{d}^2 u.
\end{aligned}$$

Proof. Let M be the matrix on the right-hand side of the equation. For any $A \in \mathcal{A}$, we have

$$\begin{aligned}
\mathrm{tr}(M\overline{A}) &= -|det(\overline{G})| \int \mathrm{tr}(\overline{u}\,(\overline{\nabla} T(u))\overline{A})\,(I(G(u)) - I_0)\,\mathrm{d}^2 u \\
&= -|det(\overline{G})| \int \mathrm{tr}((\overline{\nabla} T(u))\overline{A}\,\overline{u})\,(I(G(u)) - I_0)\,\mathrm{d}^2 u \\
&= -|det(\overline{G})| \int (\overline{\nabla} T(u))\overline{A}\,\overline{u}\,(I(G(u)) - I_0)\,\mathrm{d}^2 u \\
&= \rho_{I,T\,*}(G)(A)
\end{aligned}$$

showing that M does indeed represent $\rho_{I,T\,*}(G)$. ∎

4.5 The mass of a symbol token

The *mass*, m, of a symbol token, whose type has template T, and which is embedded in the image I by G, is defined by

$$m = |det(\overline{G})| \int T(u)\, \mathrm{d}^2 u = \int T(G^{-1}(x))\, \mathrm{d}^2 x.$$

5. Fleximaps

5.1 Introduction

A symbol contains other symbols as parts. The geometric relation between a part and the whole, or between two neighbouring parts, is not absolutely rigid but is allowed to vary along certain degrees of freedom, by certain amounts. For example, figure 1 shows a symbol token of type 'A', with three line tokens as parts. The figure shows the *nominal* relationships between the parts, i.e., the relationships in the absence of noise. But the relationship between two parts σ_1 and σ_2 may vary in several ways: σ_1 may rotate (by a limited angle) around a certain point on σ_2; σ_1 may slide up or down σ_2 (by a limited distance); and σ_1 may be shifted across σ_2 (but only by a very small amount).

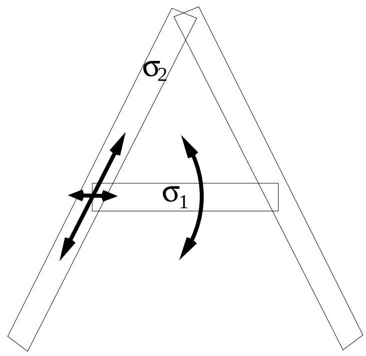

Figure 1. The geometric relation between symbol tokens σ_1 and σ_2 may vary along several degrees of freedom.

The purpose of a fleximap is to represent both the nominal geometric relationship and the permitted degrees of variation. The actual geometric relationship G between the symbol tokens σ_1 and σ_2 may be expressed in the form $G = F \cdot \exp A$, where F is the nominal relationship and A is an element of the Lie algebra $\mathcal{A}$, representing the deviation of the actual relationship from the nominal one.

We wish to impose 'soft' constraints on the deviation A. For example, we would like to say, not 'you may rotate up to $\pm\theta_0$ radians and no further', but 'there is a penalty of $k\theta^2$ for a rotation by θ radians'; the smaller the value of the coefficient k, the more rotation is tolerated. In general, the penalty function is a quadratic function depending on six independent variables (e.g., angle of rotation around a certain point, distance of translation in a certain direction, amount of dilation about a certain centre, etc.); there must always be six degrees of freedom because $\mathcal{A}$ is a six-dimensional vector space. This quadratic penalty function is technically known as a *metric tensor*.

The fleximap, then, is a pair (F,g), where F is the nominal relationship and g is the metric tensor specifying the penalty for deviations from the nominal.

5.2 Metric tensors

A *metric tensor* is a function $g: \mathcal{A} \to (\mathcal{A} \to \mathbf{R})$ such that

- $\forall c_1, c_2 \in \mathbf{R}\ \forall A, B_1, B_2 \in \mathcal{A}\ g(A)(c_1 B_1 + c_2 B_2) = c_1 g(A)(B_1) + c_2 g(A)(B_2)$,

- $\forall A, B \in \mathcal{A}\ g(A)(B) = g(B)(A)$,
- $\forall A \in \mathcal{A}\ A \neq 0 \Rightarrow g(A)(A) > 0$.

Note that, as a consequence of the first condition, $\forall A \in \mathcal{A}\ g(A) \in \mathcal{F}$; in fact, g is an isomorphism from $\mathcal{A}$ to $\mathcal{F}$.

5.3 Fleximaps

A *fleximap* is a pair (F, g), where $F \in \mathcal{G}$ and g is a metric tensor. Let *Flex* be the set of all fleximaps. Associated with any fleximap $\tau = (F, g)$ is an *energy function* $E_\tau : \mathcal{G} \to \mathbf{R}$ defined by

$$\forall G \in \mathcal{G} \quad E_\tau(G) = g(\log(F^{-1} \cdot G))(\log(F^{-1} \cdot G)).$$

The informal interpretation of a fleximap $\tau = (F, g)$ is as indicated above. Any affine transformation G not too far from F can be represented in the form $G = F \cdot \exp A$, for some $A \in \mathcal{A}$; then we have $E_\tau(G) = g(A)(A)$. F is called the *nominal* transformation, A measures how far G deviates from F, and $E_\tau(G)$ is the penalty incurred by the deviation. The way in which this works can be seen most clearly if we adopt a basis for $\mathcal{A}$ under which g is diagonal: then, in terms of components, we have $g(A)(A) = \sum_{i=1}^{6} g_{ii}(A^i)^2$. For example, the six basis elements might represent a rotation around a certain point, a dilation about a certain centre, a one-dimensional stretch in a certain direction, a shear in a certain direction, a translation in a certain direction, and a translation in another direction; these types of deviation would be penalised at the rates $g_{11}, g_{22}, g_{33}, g_{44}, g_{55}, g_{66}$ respectively. A metric tensor is a convenient way of summing up the six independent degrees of freedom and the six penalty coefficients $g_{11}, g_{22}, g_{33}, g_{44}, g_{55}, g_{66}$ associated with them.

The derivative of E_τ may be calculated as follows. For any $G \in \mathcal{G}$ and any $V \in \mathcal{A}$, we have

$$\log(F^{-1} \cdot G \cdot \exp V) = \log(F^{-1} \cdot G) + \log_*(F^{-1} \cdot G)(V) + o(V)$$

(see §3.10), so

$$\begin{aligned} E_\tau(G \cdot \exp V) &= g(\log(F^{-1} \cdot G \cdot \exp V))(\log(F^{-1} \cdot G \cdot \exp V)) \\ &= g(\log(F^{-1} \cdot G))(\log(F^{-1} \cdot G)) + 2g(\log(F^{-1} \cdot G))(\log_*(F^{-1} \cdot G)(V)) + o(V) \\ &= E_\tau(G) + 2\log^*(F^{-1} \cdot G)(g(\log(F^{-1} \cdot G)))(V) + o(V) \end{aligned}$$

so $E_{\tau *}(G)(V) = 2\log^*(F^{-1} \cdot G)(g(\log(F^{-1} \cdot G)))$.

Given a fleximap (F, g) and a transformation $G \in \mathcal{G}$, we can define the composition of the two by

$$\begin{aligned} G \cdot (F, g) &= (G \cdot F, g) \\ (F, g) \cdot G &= (F \cdot G, g') \end{aligned}$$

where g' is the metric defined by

$$\forall A, B \in \mathcal{A} \quad g'(A)(B) = g(Ad(G)(A))(Ad(G)(B)).$$

LEMMA 1. For any fleximap τ and any $G, G' \in \mathcal{G}$,

$$E_{G' \cdot \tau}(G' \cdot G) = E_\tau(G) = E_{\tau \cdot G'}(G \cdot G').$$

Proof. Let $\tau = (F,g)$. For the first equation,

$$\begin{aligned} E_{G'\cdot\tau}(G'\cdot G) &= g(\log((G'\cdot F)^{-1}\cdot G'\cdot G))(\log((G'\cdot F)^{-1}\cdot G'\cdot G)) \\ &= g(\log(F^{-1}\cdot G))(\log(F^{-1}\cdot G)) \\ &= E_\tau(G). \end{aligned}$$

For the second equation,

$$\begin{aligned} E_{\tau\cdot G'}(G\cdot G') &= g'(\log((F\cdot G')^{-1}\cdot G\cdot G'))(\log((F\cdot G')^{-1}\cdot G\cdot G')) \\ &= g'(\log(G'^{-1}\cdot F^{-1}\cdot G\cdot G'))(\log(G'^{-1}\cdot F^{-1}\cdot G\cdot G')) \\ &= g'(Ad(G'^{-1})(\log(F^{-1}\cdot G)))(Ad(G'^{-1})(\log(F^{-1}\cdot G))) \\ &= g(\log(F^{-1}\cdot G))(\log(F^{-1}\cdot G)) = E_\tau(G) \end{aligned}$$

using lemma 3.3. ∎

6. Networks and Homomorphisms

6.1 Introduction

The central concept of this paper is that of a *network*, which represents a symbol system. Networks are used for two purposes: a grammar is represented as a network of symbol types, and a particular pattern is represented as a network of symbol tokens. The relationship between a pattern and a grammar is represented by a *homomorphism* from the pattern network to the grammar network. The task of constructing the homomorphism is called *parsing*. (All this is very different from conventional grammatical formalisms, in which a grammar is a set of production rules and the relationship between a pattern and a grammar is represented by a derivation or a parse tree.)

A pattern is the system of all symbol tokens present (or believed to be present) in a given image. The position of each symbol token in the image is expressed by its *embedding*, an affine transformation. The embeddings of all the symbol tokens of the pattern are combined together into a single mathematical object, called the *embedding token* of the pattern.

The grammar includes constraints on the embedding token, expressed in terms of fleximaps. These constraints are combined together into a single mathematical object, called the *embedding type* of the grammar.

6.2 Networks

A *network* is a 9-tuple $(\Sigma, N, H, E, W, P, A, F, S)$, where Σ, N, H, E are disjoint sets, and $W, P{:}N \to \Sigma$, $A{:}H \to N$ and $F, S{:}E \to H$. The elements of Σ, N, H, E are called *symbols*, *nodes*, *hooks* and *edges*, respectively. The purpose of a node is to represent a part-whole relation between symbols: if a symbol σ_1 is a part of a symbol σ_2 then there is a node n with $P(n) = \sigma_1$ and $W(n) = \sigma_2$. Each node has zero or more hooks attached to it; if a hook h is attached to a node n then $A(h) = n$. The purpose of an edge is to represent neighbourhood relations between two parts of a common whole: if symbols σ_1 and σ_2 are both parts of a symbol σ_3, and n_1 and n_2 are the nodes representing the two part-whole relations, then there may be an edge e from one of the hooks of n_1 to one of the hooks of n_2; this will be represented by $F(e) = h_1$ and $S(e) = h_2$, where h_1 is a hook of n_1 and h_2 is a hook of n_2.

A network is *coherent* iff W is the coequaliser of $A \circ F$ and $A \circ S$ in the category of sets. This is equivalent to the conjunction of the following two conditions:

- $W \circ A \circ F = W \circ A \circ S$,
- for every $\sigma \in \Sigma$, the graph whose set of nodes is $W^{-1}(\{\sigma\})$ and whose set of edges is $F^{-1}(A^{-1}(W^{-1}(\{\sigma\})))$, where each edge e connects $A(F(e))$ to $A(S(e))$, is connected.

A network is *definite* iff the following conditions hold:

- $\forall h \in H \ |F^{-1}(\{h\})| + |S^{-1}(\{h\})| = 1$,
- $\forall \sigma \in \Sigma \ |P^{-1}(\{\sigma\})| \leq 1$.

A grammar is required to be coherent but not necessarily definite. A pattern is required to be both coherent and definite at the end of the recognition process; but while it is being constructed it is not required to be coherent or definite.

6.3 Homomorphisms

The parse of a pattern is represented by a homomorphism between the pattern and the grammar. If $\mathcal{N} = (\Sigma, N, H, E, W, P, A, F, S)$ and $\mathcal{N}' = (\Sigma', N', H', E', W', P', A', F', S')$ are networks, a *homomorphism* $f{:}\,\mathcal{N} \to \mathcal{N}'$ is a function from $\Sigma \cup N \cup H \cup E$ to $\Sigma' \cup N' \cup H' \cup E'$ such that

$$f|_\Sigma{:}\,\Sigma \to \Sigma', \quad f|_N{:}\,N \to N', \quad f|_H{:}\,H \to H', \quad f|_E{:}\,E \to E',$$
$$W' \circ f = f \circ W, \quad P' \circ f = f \circ P, \quad F' \circ f = f \circ F, \quad S' \circ f = f \circ S, \quad \text{and}$$

$$\begin{array}{ccc} H & \xrightarrow{f|_H} & H' \\ \big\downarrow A & & \big\downarrow A' \\ N & \xrightarrow{f|_N} & N' \end{array} \quad \text{is a pullback (in the category of sets).}$$

(Note that the pullback condition means that, for every $n \in N$, f maps the hooks of n bijectively onto the hooks of $f(n)$.)

An *isomorphism* is a homomorphism that is a bijection. An *automorphism* is an isomorphism from a network $\mathcal{N}$ to itself.

6.4 Embeddings

An *embedding token* for a network $(\Sigma, N, H, E, W, P, A, F, S)$ is a function $u{:}\,\Sigma \cup N \to \mathcal{G}$. It specifies the way a pattern is embedded in the image plane: $u(\sigma)$ is the embedding of a symbol σ, and $u(n)$ is the embedding of a node n. If $P(n) = \sigma$ then $u(\sigma)$ and $u(n)$ will be very closely related: they will differ only by one of σ's symmetry transformations (this constraint is called the *symmetry condition*; see §7.3 below).

Given two embedding tokens, u_1 and u_2, for the same network, we can define a new embedding token $u_1 \cdot u_2$ by

$$\forall x \in \Sigma \cup N \quad (u_1 \cdot u_2)(x) = u_1(x) \cdot u_2(x).$$

Given a homomorphism $f{:}\,\mathcal{N} \to \mathcal{N}'$ and an embedding token u for $\mathcal{N}'$, clearly $u \circ f$ is an embedding token for $\mathcal{N}$.

An *embedding type* for a network $(\Sigma, N, H, E, W, P, A, F, S)$ is a quintuple $(\mathit{con}, \mathit{rel}, \mathit{symm}, \mathit{tem}, \mathit{in})$, in which $\mathit{con}{:}\,N \to \mathit{Flex}$, $\mathit{rel}{:}\,E \to \mathit{Flex}$, $\mathit{symm}{:}\,\Sigma \to \mathit{sub}(\mathcal{G})$, $\mathit{tem}{:}\,\Sigma \to \mathit{Tem}$, and $\mathit{in}{:}\,\Sigma \to \mathit{Flex}$ where Flex is the set of all fleximaps, $\mathit{sub}(\mathcal{G})$ is the set of all subgroups of $\mathcal{G}$, and Tem is the set of all templates; for every $\sigma \in \Sigma$, the fleximap $\mathit{in}(\sigma)$ must have nominal part equal to the identity transformation.

The purpose of an embedding type is to impose constraints on the embedding token. For any node n, $\mathit{con}(n)$ is a fleximap describing the relationship between the embeddings of a part and a whole (or, more precisely, of the node n and the symbol $W(n)$). For any edge e, $\mathit{rel}(e)$ is a fleximap describing the relationship between the embeddings of two neighbouring parts (or, more precisely, of their nodes $A(F(e))$ and $A(S(e))$). For any symbol σ, $\mathit{symm}(\sigma)$ is its symmetry group, $\mathit{tem}(\sigma)$ is its template, and $\mathit{in}(\sigma)$ is (I, g), where I is the identity transformation and g is called the *inertial* metric of σ and determines how σ moves in response to forces (see §7.7).

The *Match* function, defined below in §7.3, defines how well a given embedding token matches a given embedding type.

Given an embedding type $v = (con_1, rel_1, symm_1, tem_1, in_1)$ and an embedding token u for a network $(\Sigma, N, H, E, W, P, A, F, S)$, we can define an embedding type $v \cdot u$ for the same network by $v \cdot u = (con_2, rel_2, symm_1, tem_2, in_2)$, where

$$\begin{aligned}
\forall n \in N \quad con_2(n) &= u(W(n))^{-1} \cdot con_1(n) \cdot u(n) \\
\forall e \in E \quad rel_2(e) &= u(A(S(e)))^{-1} \cdot rel_1(e) \cdot u(A(F(e))) \\
\forall \sigma \in \Sigma \quad tem_2(\sigma) &= tem_1(\sigma) \circ u(\sigma) \\
\forall \sigma \in \Sigma \quad in_2(\sigma) &= u(\sigma)^{-1} \cdot in_1(\sigma) \cdot u(\sigma)
\end{aligned}$$

Given an embedding type $v = (con, rel, symm, tem, in)$ for a network $\mathcal{N}'$ and a homomorphism $f{:}\mathcal{N} \to \mathcal{N}'$, we can define an embedding type $v \circ f$ for $\mathcal{N}$ by

$$v \circ f = (con \circ f, rel \circ f, symm \circ f, tem \circ f, in \circ f).$$

LEMMA 1. If $f{:}\mathcal{N} \to \mathcal{N}'$ is a homomorphism, v is an embedding type for $\mathcal{N}'$, and u is an embedding token for $\mathcal{N}'$, then

$$(v \cdot u) \circ f = (v \circ f) \cdot (u \circ f).$$

Proof. Let $\mathcal{N} = (\Sigma, N, H, E, W, P, A, F, S)$, $\mathcal{N}' = (\Sigma', N', H', E', W', P', A', F', S')$ and $v = (con_1, rel_1, symm_1, tem_1, in_1)$. Then $v \cdot u = (con_2, rel_2, symm_1, tem_2, in_2)$, where

$$\begin{aligned}
\forall n \in N' \quad con_2(n) &= u(W'(n))^{-1} \cdot con_1(n) \cdot u(n) \\
\forall e \in E' \quad rel_2(e) &= u(A'(S'(e)))^{-1} \cdot rel_1(e) \cdot u(A'(F'(e))) \\
\forall \sigma \in \Sigma' \quad tem_2(\sigma) &= tem_1(\sigma) \circ u(\sigma) \\
\forall \sigma \in \Sigma' \quad in_2(\sigma) &= u(\sigma)^{-1} \cdot in_1(\sigma) \cdot u(\sigma)
\end{aligned}$$

and hence

$$(v \cdot u) \circ f = (con_2 \circ f, rel_2 \circ f, symm_1 \circ f, tem_2 \circ f, in_2 \circ f).$$

We also have $v \circ f = (con_1 \circ f, rel_1 \circ f, symm_1 \circ f, tem_1 \circ f, in_1 \circ f)$, and hence $(v \circ f) \cdot (u \circ f) = (con_3, rel_3, symm_1 \circ f, tem_3, in_3)$, where

$$\begin{aligned}
\forall n \in N \quad con_3(n) &= (u \circ f)(W(n))^{-1} \cdot (con_1 \circ f)(n) \cdot (u \circ f)(n) \\
&= u(f(W(n)))^{-1} \cdot con_1(f(n)) \cdot u(f(n)) \\
&= u(W'(f(n)))^{-1} \cdot con_1(f(n)) \cdot u(f(n)) \\
&= con_2(f(n)) = (con_2 \circ f)(n) \\
\forall e \in E \quad rel_3(e) &= (u \circ f)(A(S(e)))^{-1} \cdot (rel_1 \circ f)(e) \cdot (u \circ f)(A(F(e))) \\
&= u(f(A(S(e))))^{-1} \cdot rel_1(f(e)) \cdot u(f(A(F(e)))) \\
&= u(A'(S'(f(e))))^{-1} \cdot rel_1(f(e)) \cdot u(A'(F'(f(e)))) \\
&= rel_2(f(e)) = (rel_2 \circ f)(e) \\
\forall \sigma \in \Sigma \quad tem_3(\sigma) &= (tem_1 \circ f)(\sigma) \circ (u \circ f)(\sigma) \\
&= tem_1(f(\sigma)) \circ u(f(\sigma)) \\
&= tem_2(f(\sigma)) = (tem_2 \circ f)(\sigma) \\
\forall \sigma \in \Sigma \quad in_3(\sigma) &= (u \circ f)(\sigma)^{-1} \cdot (in_1 \circ f)(\sigma) \cdot (u \circ f)(\sigma) \\
&= u(f(\sigma))^{-1} \cdot in_1(f(\sigma)) \cdot u(f(\sigma)) \\
&= in_2(f(\sigma)) = (in_2 \circ f)(\sigma)
\end{aligned}$$

which shows that $(v \cdot u) \circ f = (v \circ f) \cdot (u \circ f)$. ∎

7. The Recognition Problem

7.1 Introduction

Given an image, we wish to detect all the symbol tokens present in it, determine how they are embedded in it and how they are related to one another, and relate the system of symbol tokens to the grammar. This is the *recognition problem*. To restate this using the terminology of §6: we need to construct the pattern, determine the embedding token of the pattern, and construct the homomorphism from the pattern to the grammar. The recognition problem will be stated formally in §7.4.

While the pattern is being constructed it may temporarily be in a inconsistent state: it may be incoherent or indefinite (these terms were defined in §6.2). Also, the symbols, nodes and edges of the pattern may be only 'tentatively present': this is represented by *inclusion functions* (e.g., $i(\sigma)$ is the degree to which it is believed that a symbol token σ should be present in the pattern). However, by the end of the recognition process the pattern must become coherent and definite, and every symbol, node and edge must be unequivocally present.

The recognition process is governed by a *Match* function, which measures how well a pattern matches the given image and grammar. The aim of recognition is to maximise this function. Only one aspect of the recognition problem is solved in this paper: for a given pattern and homomorphism, the optimal embedding token for the pattern is found by a process of gradient descent on the *Match* function (see §7.7).

7.2 Inclusion functions

During the course of recognition, the pattern is accompanied by *inclusion functions* $i\colon \Sigma \cup N \cup E \to [0,1]$ and $j\colon \Sigma \cup N \cup H \to [0,1]$, which are subject to the conditions

$$\begin{aligned}
\forall \sigma \in \Sigma \quad i(\sigma) &= j(\sigma) + \sum_{n \in P^{-1}(\{\sigma\})} i(n) \\
\forall n \in N \quad i(W(n)) &= j(n) + i(n) \\
\forall h \in H \quad i(A(h)) &= j(h) + \sum_{e \in F^{-1}(\{h\}) \cup S^{-1}(\{h\})} i(e)
\end{aligned}$$

These functions reflect the uncertainty during recognition concerning which parts of the pattern should really be present: $i(\sigma)$ represents the degree of belief that σ should be present, and similarly for $i(n)$ and $i(e)$ (there is no need for an inclusion value for hooks because a hook is present iff its node is). The j function is concerned with missing parts: $j(h)$ represents the degree of belief that the hook h should be present but that the correct edge to be connected to it has not yet been found; $j(n)$ represents the degree of belief that the symbol $W(n)$ should be present but the node n should not be; $j(\sigma)$ represents the degree of belief that σ should be present but as a top-level symbol, i.e., with $P^{-1}(\{\sigma\}) = \emptyset$.

At the end of recognition we must have $\forall x \in \Sigma \cup N \cup E\ i(x) = 1$, i.e., complete certainty.

In addition there is a function $B\colon H \to \mathbf{R}$ that imposes a penalty on bare hooks: if a hook h has no edges attached to it then a penalty of $B(h)$ is incurred. When a node is first formed its hooks have low values of $B(h)$, since they cannot be expected to have acquired edges yet; but if they remain bare the value of $B(h)$ is increased to force them to acquire edges. At the end of the recognition process every hook must have an edge, to satisfy the definiteness condition.

7.3 The match function between an embedding type and an embedding token

Given a network $\mathcal{N} = (\Sigma, N, H, E, W, P, A, F, S)$, an image I, an embedding token u for $\mathcal{N}$, and an embedding type $v = (con, rel, symm, tem)$ for $\mathcal{N}$, the *match function*, which measures how well u matches v and I, is defined by

$$
\begin{aligned}
&Match(I,\mathcal{N},u,v) = \\
&\sum_{\sigma\in\Sigma}\big(i(\sigma)\rho_{I,tem(\sigma)}(u(\sigma)) - j(\sigma)\theta\big) - \sum_{n\in N} i(n)E_{con(n)}(u(W(n))^{-1}\cdot u(n)) - \sum_{h\in H} j(h)B(h)\\
&-\sum_{e\in E} i(e)\big(E_{rel(e)}(u(A(S(e)))^{-1}\cdot u(A(F(e)))) + E_{in(W(A(F(e))))}(u(W(A(S(e))))^{-1}\cdot u(W(A(F(e)))))\big)
\end{aligned}
$$

where θ is a positive real constant. The *symmetry condition* for $\mathcal{N}, u, v$ is

$$\forall n\in N \quad u(P(n))^{-1}\cdot u(n) \in symm(P(n)).$$

The terms in the definition of *Match* can be explained as follows. For any symbol σ, $u(\sigma)$ is the embedding of σ in the image plane. For any node n, $u(n)$ is equal to $u(P(n))$, the embedding of the symbol $P(n)$, possibly adjusted by a symmetry transformation (a member of $symm(P(n))$). The term $\rho_{I,tem(\sigma)}(u(\sigma))$ measures how well the template of σ, embedded in the image by $u(\sigma)$, matches the image (see §4). The term θ is a fixed penalty incurred by any top-level symbol, i.e., any symbol that is not a part of any other symbol. The term $E_{con(n)}(u(W(n))^{-1}\cdot u(n))$ measures how well the part-whole relationship between $P(n)$ and $W(n)$ matches the fleximap $con(n)$. The term $B(h)$ is a penalty incurred by a hook h with no incident edges (see §7.2). The term $E_{rel(e)}(u(A(S(e)))^{-1}\cdot u(A(F(e))))$ measures how well the neighbourhood relation between the symbols $P(A(F(e)))$ and $P(A(S(e)))$ matches the fleximap $rel(e)$. The final term, $E_{in(W(A(F(e))))}(u(W(A(S(e))))^{-1}\cdot u(W(A(F(e)))))$ is only non-zero if the network is not coherent: if an edge e has $W(A(S(e))) \neq W(A(F(e)))$ then this term penalises the difference in embedding between $W(A(S(e)))$ and $W(A(F(e)))$. By the end of the recognition process, the symbols $W(A(S(e)))$ and $W(A(F(e)))$ will either have been forced together by this term and merged, or the edge will have been removed.

All these terms are weighted by inclusion functions, $i(\sigma), j(\sigma), i(n), j(h), i(e)$, to reflect the different degrees to which σ, etc., are believed to be present.

LEMMA 1. For any image I, network $\mathcal{N}$, embedding tokens u, u' for $\mathcal{N}$, and embedding type v for $\mathcal{N}$,

$$Match(I,\mathcal{N},u\cdot u', v\cdot u') = Match(I,\mathcal{N},u,v),$$

provided $u'\circ W\circ A\circ F = u'\circ W\circ A\circ S$.

Proof. Let $v = (con_1, rel_1, symm_1, tem_1, in_1)$; then $v \cdot u' = (con_2, rel_2, symm_1, tem_2, in_2)$, where

$$\begin{aligned} \forall n \in N \quad con_2(n) &= u'(W(n))^{-1} \cdot con_1(n) \cdot u'(n) \\ \forall e \in E \quad rel_2(e) &= u'(A(S(e)))^{-1} \cdot rel_1(e) \cdot u'(A(F(e))) \\ \forall \sigma \in \Sigma \quad tem_2(\sigma) &= tem_1(\sigma) \circ u'(\sigma) \\ \forall \sigma \in \Sigma \quad in_2(\sigma) &= u'(\sigma)^{-1} \cdot in_1(\sigma) \cdot u'(\sigma) \end{aligned}$$

Now, first, for any $\sigma \in \Sigma$,

$$\rho_{I,tem_2(\sigma)}((u \cdot u')(\sigma)) = \rho_{I,tem_1(\sigma) \circ u'(\sigma)}(u(\sigma) \cdot u'(\sigma)) = \rho_{I,tem_1(\sigma)}(u(\sigma))$$

using lemma 4.1.

Secondly, for any $n \in N$,

$$\begin{aligned} &E_{con_2(n)}((u \cdot u')(W(n))^{-1} \cdot (u \cdot u')(n)) \\ &\qquad = E_{u'(W(n))^{-1} \cdot con_1(n) \cdot u'(n)}(u'(W(n))^{-1} \cdot u(W(n))^{-1} \cdot u(n) \cdot u'(n)) \\ &\qquad = E_{con_1(n)}(u(W(n))^{-1} \cdot u(n)) \end{aligned}$$

by lemma 5.1.

Thirdly, for any $e \in E$, using the abbreviations $n_1 = A(F(e))$ and $n_2 = A(S(e))$,

$$\begin{aligned} E_{rel_2(e)}((u \cdot u')(n_2)^{-1} \cdot (u \cdot u')(n_1)) &= E_{u'(n_2)^{-1} \cdot rel_1(e) \cdot u'(n_1)}(u'(n_2)^{-1} \cdot u(n_2)^{-1} \cdot u(n_1) \cdot u'(n_1)) \\ &= E_{rel_1(e)}(u(n_2)^{-1} \cdot u(n_1)) \end{aligned}$$

and, using the abbreviations $\sigma_1 = W(A(F(e)))$ and $\sigma_2 = W(A(S(e)))$,

$$\begin{aligned} E_{in_2(\sigma_1)}((u \cdot u')(\sigma_2)^{-1} \cdot (u \cdot u')(\sigma_1)) &= E_{u'(\sigma_1)^{-1} \cdot in_1(\sigma_1) \cdot u'(\sigma_1)}(u'(\sigma_2)^{-1} \cdot u(\sigma_2)^{-1} \cdot u(\sigma_1) \cdot u'(\sigma_1)) \\ &= E_{in_1(\sigma_1)}(u(\sigma_2)^{-1} \cdot u(\sigma_1)) \end{aligned}$$

since $u'(\sigma_1) = u'(\sigma_2)$.

Thus each term of $Match(I, \mathcal{N}, u \cdot u', v \cdot u')$ equals the corresponding term of $Match(I, \mathcal{N}, u, v)$. ∎

7.4 The recognition problem

Given a coherent network $\mathcal{N}_0$ (representing a grammar), an embedding type v for $\mathcal{N}_0$, and an image I, the *recognition problem* is to find a coherent and definite network $\mathcal{N}_1$ (representing a pattern), with inclusion function i satisfying $\forall x \in \Sigma_1 \cup N_1 \cup E_1 \; i(x) = 1$, a homomorphism $p{:}\, \mathcal{N}_1 \to \mathcal{N}_0$ (representing a parse of the pattern according to the grammar), and an embedding token u for $\mathcal{N}_1$, that maximises $Match(I, \mathcal{N}_1, u, v \circ p)$, subject to the symmetry condition for $\mathcal{N}_1, u, v \circ p$.

7.5 Symmetries

A *symmetry* of a network $\mathcal{N} = (\Sigma, N, H, E, W, P, A, F, S)$, with respect to the embedding type $v = (con, rel, symm, tem, in)$, is an automorphism $a: \mathcal{N} \to \mathcal{N}$ together with an embedding token s for $\mathcal{N}$, such that

$\forall \sigma \in \Sigma \; s(\sigma) \in symm(\sigma)$,
$\forall n \in N \; s(n) \in symm(P(n))$,
$v \cdot s = v \circ a$.

LEMMA 2. (Global application of a symmetry.) Given a symmetry a, s of the grammar $\mathcal{N}_0$ with respect to the embedding type v, a homomorphism $p: \mathcal{N}_1 \to \mathcal{N}_0$, and an embedding token u for $\mathcal{N}_1$, we can produce a new parse $p' = a \circ p: \mathcal{N}_1 \to \mathcal{N}_0$ and embedding token $u' = u \cdot (s \circ p)$. Provided the grammar is coherent, this gives

$$Match(I, \mathcal{N}_1, u', v \circ p') = Match(I, \mathcal{N}_1, u, v \circ p),$$

and the symmetry condition holds for $\mathcal{N}_1, u', v \circ p'$ iff it holds for $\mathcal{N}_1, u, v \circ p$.

Proof. Let $\mathcal{N}_0 = (\Sigma_0, N_0, H_0, E_0, W_0, P_0, A_0, F_0, S_0)$ and $\mathcal{N}_1 = (\Sigma_1, N_1, H_1, E_1, W_1, P_1, A_1, F_1, S_1)$. Applying the definition of a symmetry, followed by lemma 6.1 and lemma 1, gives

$$\begin{aligned} Match(I, \mathcal{N}_1, u', v \circ p') &= Match(I, \mathcal{N}_1, u \cdot (s \circ p), v \circ a \circ p) \\ &= Match(I, \mathcal{N}_1, u \cdot (s \circ p), (v \cdot s) \circ p) \\ &= Match(I, \mathcal{N}_1, u \cdot (s \circ p), (v \circ p) \cdot (s \circ p)) \\ &= Match(I, \mathcal{N}_1, u, v \circ p). \end{aligned}$$

Note that the application of lemma 1 is justified since we have $s \circ p \circ W_1 \circ A_1 \circ F_1 = s \circ W_0 \circ A_0 \circ F_0 \circ p = s \circ W_0 \circ A_0 \circ S_0 \circ p = s \circ p \circ W_1 \circ A_1 \circ S_1$, since $\mathcal{N}_0$ is coherent.

The symmetry condition is checked as follows. Let $v = (con, rel, symm, tem, in)$. For any $n \in N_1$, we have

$$u'(P_1(n))^{-1} \cdot u'(n) = s(p(P_1(n)))^{-1} \cdot u(P_1(n))^{-1} \cdot u(n) \cdot s(p(n)).$$

Note that, since a, s is a symmetry of $\mathcal{N}_0$, we have $s(p(P_1(n))) \in symm(p(P_1(n)))$ and $s(p(n)) \in symm(P_0(p(n))) = symm(p(P_1(n)))$, and $symm = symm \circ a$. Now, the symmetry condition for $\mathcal{N}_1, u, v \circ p$ requires

$$u(P_1(n))^{-1} \cdot u(n) \in symm(p(P_1(n))),$$

whereas the symmetry condition for $\mathcal{N}_1, u', v \circ p'$ requires

$$u'(P_1(n))^{-1} \cdot u'(n) \in symm(a(p(P_1(n)))).$$

From the above information we see that these two conditions are equivalent. ∎

7.6 The derivative of the match function

Given a network $\mathcal{N} = (\Sigma, N, H, E, W, P, A, F, S)$, an image I, an embedding token u for $\mathcal{N}$, and an embedding type $v = (con, rel, symm, tem, in)$ for $\mathcal{N}$, we can consider the derivative of the *Match* function with respect to the embeddings of the symbols: that is, we can consider the change in $Match(I, \mathcal{N}, u, v)$ produced by a small change in u.

Let us introduce the following abbreviations:

$$\begin{aligned}
\forall n \in N \quad S_n &= u(P(n))^{-1} \cdot u(n) \\
\forall n \in N \quad C_n &= u(W(n))^{-1} \cdot u(n) \\
\forall e \in E \quad R_e &= u(A(S(e)))^{-1} \cdot u(A(F(e))) \\
\forall e \in E \quad R'_e &= u(W(A(S(e))))^{-1} \cdot u(W(A(F(e))))
\end{aligned}$$

This gives

$$\begin{aligned}
Match(I, \mathcal{N}, u, v) = &\sum_{\sigma \in \Sigma} \big(i(\sigma) \rho_{I, tem(\sigma)}(u(\sigma)) - j(\sigma)\theta \big) - \sum_{n \in N} i(n) E_{con(n)}(C_n) - \sum_{h \in H} j(h) B(h) \\
&- \sum_{e \in E} i(e) \big(E_{rel(e)}(R_e) + E_{in(W(A(F(e))))}(R'_e) \big)
\end{aligned}$$

and the symmetry condition is

$$\forall n \in N \quad S_n \in symm(P(n)).$$

Now, consider the effect of making a small change in the embedding token from u to $u \cdot \Delta u$, where

$$\begin{aligned}
\forall \sigma \in \Sigma \quad \Delta u(\sigma) &= \exp(\varepsilon V_\sigma) \\
\forall n \in N \quad \Delta u(n) &= \exp(\varepsilon V_n)
\end{aligned}$$

where ε is a real number in a neighbourhood of 0 and $V_\sigma, V_n \in \mathcal{A}$. The movement V_σ of each symbol σ may be chosen arbitrarily, but, in order to preserve the symmetry condition, the movement V_n of each node n must then be $V_n = Ad(S_n^{-1})(V_{P(n)})$, thus leaving S_n invariant.

THEOREM 3. For Δu defined as above,

$$Match(I, \mathcal{N}, u \cdot \Delta u, v) = Match(I, \mathcal{N}, u, v) + \varepsilon \sum_{\sigma \in \Sigma} F_\sigma(V_\sigma) + o(\varepsilon)$$

where, for each $\sigma \in \Sigma$, $n \in N$ and $e \in E$,

$$\begin{aligned}
F_\sigma &= F_\sigma^0 + F_\sigma^1 + F_\sigma^2 + F_\sigma^3 \in \mathcal{F} \\
F_\sigma^0 &= i(\sigma)\, \rho_{I, tem(\sigma)\,*}(u(\sigma)) \\
F_\sigma^1 &= -\sum_{n \in W^{-1}(\{\sigma\})} Ad(C_n^{-1})^\dagger(F_n) \quad + \sum_{n \in P^{-1}(\{\sigma\})} Ad(S_n^{-1})^\dagger(F_n) \\
F_\sigma^2 &= \sum_{n \in P^{-1}(\{\sigma\})} Ad(S_n^{-1})^\dagger \Big(-\sum_{e \in S^{-1}(A^{-1}(\{n\}))} Ad(R_e^{-1})^\dagger(F_e) \quad + \sum_{e \in F^{-1}(A^{-1}(\{n\}))} F_e \Big) \\
F_\sigma^3 &= -\sum_{e \in S^{-1}(A^{-1}(W^{-1}(\{\sigma\})))} Ad({R'_e}^{-1})^\dagger(F'_e) \quad + \sum_{e \in F^{-1}(A^{-1}(W^{-1}(\{\sigma\})))} F'_e \\
F_n &= -i(n)\, E_{con(n)\,*}(C_n) \\
F_e &= -i(e)\, E_{rel(e)\,*}(R_e) \\
F'_e &= -i(e)\, E_{in(W(A(F(e))))\,*}(R'_e)
\end{aligned}$$

Proof. The effect of the change $u \mapsto u \cdot \Delta u$ on C_n is

$$\begin{aligned} C_n &\mapsto \exp(-\varepsilon V_{W(n)}) \cdot C_n \cdot \exp(\varepsilon V_n) \\ &= C_n \cdot \exp(-\varepsilon Ad(C_n^{-1})(V_{W(n)})) \cdot \exp(\varepsilon V_n) \\ &= C_n \cdot \exp\left(-\varepsilon Ad(C_n^{-1})(V_{W(n)}) + \varepsilon V_n + o(\varepsilon)\right) \end{aligned}$$

using lemma 3.10. The effect of the change $u \mapsto u \cdot \Delta u$ on R_e is

$$\begin{aligned} R_e &\mapsto \exp(-\varepsilon V_{A(S(e))}) \cdot R_e \cdot \exp(\varepsilon V_{A(F(e))}) \\ &= R_e \cdot \exp(-\varepsilon Ad(R_e^{-1})(V_{A(S(e))})) \cdot \exp(\varepsilon V_{A(F(e))}) \\ &= R_e \cdot \exp(\varepsilon V_e + o(\varepsilon)) \end{aligned}$$

where $V_e = -Ad(R_e^{-1})(V_{A(S(e))}) + V_{A(F(e))}$, using lemma 3.10 again. The effect of the change $u \mapsto u \cdot \Delta u$ on R'_e is

$$\begin{aligned} R'_e &\mapsto \exp(-\varepsilon V_{W(A(S(e)))}) \cdot R'_e \cdot \exp(\varepsilon V_{W(A(F(e)))}) \\ &= R'_e \cdot \exp(-\varepsilon Ad({R'_e}^{-1})(V_{W(A(S(e)))})) \cdot \exp(\varepsilon V_{W(A(F(e)))}) \\ &= R'_e \cdot \exp(\varepsilon V'_e + o(\varepsilon)) \end{aligned}$$

where $V'_e = -Ad({R'_e}^{-1})(V_{W(A(S(e)))}) + V_{W(A(F(e)))}$, by lemma 3.10. We are interested in calculating the consequent change in the value of the match function. Using lemma 3.11,

$$\begin{aligned} Match(I, \mathcal{N}, u \cdot \Delta u, v) = Match(I, \mathcal{N}, u, v) &+ \varepsilon \sum_{\sigma \in \Sigma} i(\sigma)\, \rho_{I,tem(\sigma)\,*}(u(\sigma))(V_\sigma) \\ &- \varepsilon \sum_{n \in N} i(n)\, E_{con(n)\,*}(C_n)(-Ad(C_n^{-1})(V_{W(n)}) + V_n) \\ &- \varepsilon \sum_{e \in E} i(e)\, \left(E_{rel(e)\,*}(R_e)(V_e) + E_{in(W(A(F(e))))\,*}(R'_e)(V'_e)\right) + o(\varepsilon). \quad (*) \end{aligned}$$

Using the above definition of F_n, the sum over nodes in equation $(*)$ can be rewritten:

$$\begin{aligned} -\varepsilon \sum_{n \in N} i(n)\, & E_{con(n)\,*}(C_n)(-Ad(C_n^{-1})(V_{W(n)}) + V_n) \\ &= -\varepsilon \sum_{n \in N} F_n(Ad(C_n^{-1})(V_{W(n)})) + \varepsilon \sum_{n \in N} F_n(V_n) \\ &= -\varepsilon \sum_{n \in N} F_n(Ad(C_n^{-1})(V_{W(n)})) + \varepsilon \sum_{n \in N} F_n(Ad(S_n^{-1})(V_{P(n)})) \\ &= -\varepsilon \sum_{n \in N} Ad(C_n^{-1})^\dagger(F_n)(V_{W(n)}) + \varepsilon \sum_{n \in N} Ad(S_n^{-1})^\dagger(F_n)(V_{P(n)}) \\ &= \varepsilon \sum_{\sigma \in \Sigma} F^1_\sigma(V_\sigma). \end{aligned}$$

Similarly, using the above definition of F_e, we can rewrite part of the sum over edges in equation $(*)$ as

$$\begin{aligned} -\varepsilon \sum_{e \in E} i(e)\, E_{rel(e)\,*}(R_e)(V_e) &= \varepsilon \sum_{e \in E} F_e(-Ad(R_e^{-1})(V_{A(S(e))}) + V_{A(F(e))}) \\ &= -\varepsilon \sum_{e \in E} F_e(Ad(R_e^{-1})(V_{A(S(e))})) + \varepsilon \sum_{e \in E} F_e(V_{A(F(e))}) \\ &= -\varepsilon \sum_{e \in E} F_e(Ad(R_e^{-1})(Ad(S^{-1}_{A(S(e))})(V_{P(A(S(e)))}))) + \varepsilon \sum_{e \in E} F_e(Ad(S^{-1}_{A(F(e))})(V_{P(A(F(e)))})) \\ &= -\varepsilon \sum_{e \in E} Ad(S^{-1}_{A(S(e))})^\dagger(Ad(R_e^{-1})^\dagger(F_e))(V_{P(A(S(e)))}) + \varepsilon \sum_{e \in E} Ad(S^{-1}_{A(F(e))})^\dagger(F_e)(V_{P(A(F(e)))}) \\ &= \varepsilon \sum_{\sigma \in \Sigma} F^2_\sigma(V_\sigma). \end{aligned}$$

Using the above definition of F'_e, we can rewrite the other part of the sum over edges in equation $(*)$:

$$\begin{aligned}-\varepsilon\sum_{e\in E} i(e)\,E_{in(W(A(F(e))))\,*}(R'_e)(V'_e) &= \varepsilon\sum_{e\in E} F'_e(-Ad(R'^{-1}_e)(V_{W(A(S(e)))}) + V_{W(A(F(e)))}) \\ &= -\varepsilon\sum_{e\in E} F'_e(Ad(R'^{-1}_e)(V_{W(A(S(e)))})) + \varepsilon\sum_{e\in E} F'_e(V_{W(A(F(e)))}) \\ &= -\varepsilon\sum_{e\in E} Ad(R'^{-1}_e)^{\dagger}(F'_e)(V_{W(A(S(e)))}) + \varepsilon\sum_{e\in E} F'_e(V_{W(A(F(e)))}) \\ &= \varepsilon\sum_{\sigma\in\Sigma} F^3_\sigma(V_\sigma).\end{aligned}$$

Hence equation $(*)$ finally reduces to

$$Match(I,\mathcal{N},u\cdot\Delta u,v) = Match(I,\mathcal{N},u,v) + \varepsilon\sum_{\sigma\in\Sigma} F_\sigma(V_\sigma) + o(\varepsilon)$$

as required. ∎

Thus $F_\sigma\in\mathcal{F}$ measures the effect on the value of the *Match* function, to first order in ε, of the change in the embedding of σ. F_σ is called the *force on* σ, and consists of four components: F^0_σ, the force exerted by σ's template; F^1_σ, the force exerted by the part-whole fleximaps in which σ participates; F^2_σ, the force exerted by the neighbourhood fleximaps in which σ participates; and F^3_σ, the force that pulls two symbols $W(A(F(e)))$ and $W(A(S(e)))$ together, whenever any edge e has $W(A(F(e)))\neq W(A(S(e)))$.

7.7 Gradient ascent on the match function

The aim of gradient ascent is to make a small change to the embedding token u of the pattern so as to increase $Match(I,\mathcal{N}_1,u,v\circ p)$ as much as possible (while preserving the symmetry condition). Let $\mathcal{N}_1 = (\Sigma_1,N_1,H_1,E_1,W_1,P_1,A_1,F_1,S_1)$ and $v = (con,rel,symm,tem,in)$. As shown in theorem 3, the effect of changing the embeddings of the symbol tokens by

$$\forall\sigma\in\Sigma_1\quad u(\sigma)\mapsto u(\sigma)\cdot\exp(\varepsilon V_\sigma)$$

(with a corresponding change in the embeddings of the nodes) is to produce a change in $Match(I,\mathcal{N}_1,u,v\circ p)$ of $\varepsilon\sum_{\sigma\in\Sigma_1}F_\sigma(V_\sigma)$, to first order in ε.

The cost of making this change is defined as $\frac{1}{2}\varepsilon\sum_{\sigma\in\Sigma_1} m_\sigma g_\sigma(V_\sigma)(V_\sigma)$, where m_σ is the mass of σ (see §4.5) and $(I,g_\sigma) = in(p(\sigma))$, i.e., g_σ is the inertial metric tensor for σ provided by $v\circ p$. Hence we wish to choose the V_σ's to maximise

$$\begin{aligned}&\varepsilon\sum_{\sigma\in\Sigma_1}F_\sigma(V_\sigma) - \tfrac{1}{2}\varepsilon\sum_{\sigma\in\Sigma_1} m_\sigma g_\sigma(V_\sigma)(V_\sigma) \\ &\qquad= \tfrac{1}{2}\varepsilon\sum_{\sigma\in\Sigma_1}\tfrac{1}{m_\sigma}F_\sigma(g_\sigma^{-1}(F_\sigma)) - \tfrac{1}{2}\varepsilon\sum_{\sigma\in\Sigma_1}\tfrac{1}{m_\sigma}(m_\sigma g_\sigma(V_\sigma) - F_\sigma)(m_\sigma V_\sigma - g_\sigma^{-1}(F_\sigma))\end{aligned}$$

This is maximised by taking $V_\sigma = \frac{1}{m_\sigma}g_\sigma^{-1}(F_\sigma)$, for each σ.

7.8 The grand conclusion

Given a grammar $\mathcal{N}_0$, an embedding type v for $\mathcal{N}_0$, and an image I, the recognition problem is to construct a pattern $\mathcal{N}_1$, a parse homomorphism $p{:}\mathcal{N}_1 \to \mathcal{N}_0$, and an embedding token u for $\mathcal{N}_1$. This paper has solved one part of the problem: it has shown that u can be optimised incrementally by applying the gradient ascent rule

$$\begin{aligned} &\forall \sigma \in \Sigma_1 \quad u(\sigma) \mapsto u(\sigma) \cdot \exp(\varepsilon V_\sigma) \\ &\forall n \in N_1 \quad u(n) \mapsto u(n) \cdot \exp(\varepsilon Ad(S_n^{-1})(V_{P(n)})), \end{aligned}$$

where

$$V_\sigma = \tfrac{1}{m_\sigma} g_\sigma^{-1}(F_\sigma)$$

and F_σ is the force on σ, defined in theorem 3.

References

Barnden, J.A. & Pollack, J.B. (1991) Problems for high-level connectionism. In Barnden, J.A. & Pollack, J.B. (eds) *High-Level Connectionist Models.* Advances in Connectionist and Neural Computation Theory, vol. 1. Norwood, New Jersey: Ablex, pp. 1–16.

Boothby, W.M. (1986) *An Introduction to Differentiable Manifolds and Riemannian Geometry.* Orlando, Florida: Academic Press.

Dinsmore, J. (ed.) (1992) *The Symbolic and Connectionist Paradigms: Closing the Gap.* Hillsdale, New Jersey: Lawrence Erlbaum Associates.

Fletcher, P. (2001) Connectionist learning of regular graph grammars. *Connection Science*, **13**, no. 2, 127–188.

Fodor, J.A. & Pylyshyn, Z.W. (1988) Connectionism and cognitive architecture: a critical analysis. *Cognition*, **28**, nos 1 & 2, 3–71.

Garfield, J.L. (1997) Mentalese not spoken here: computation, cognition and causation. *Philosophical Psychology*, **10**, no. 4, 413–435.

Hadley, R.F. (1999) Connectionism and novel combinations of skills: implications for cognitive architecture. *Minds and Machines*, **9**, no. 2, 197–221.

Hausner, M & Schwartz, J.T. (1968) *Lie Groups, Lie Algebras.* London: Nelson.

Horgan, T. & Tienson, J. (1996) *Connectionism and the Philosophy of Psychology.* Cambridge, Massachusetts: MIT Press.

Price, J.F. (1977) *Lie Groups and Compact Groups.* London Mathematical Society Lecture Note Series, vol. 25. Cambridge: Cambridge University Press.

Smolensky, P. (1988) On the proper treatment of connectionism. *Behavioral and Brain Sciences*, **11**, no. 1, 1–23.

Sougné, J. (1998) Connectionism and the problem of multiple instantiation. *Trends in Cognitive Science*, **2**, no. 5, 183–189.