Network: Computation in Neural Systems **5**(4), November 1994

Self-organisation of transformation-invariant detectors for constituents of perceptual patterns

Chris J.S. Webber

Departments of Physiology, Engineering & Physics, University of Cambridge
Now at: Pattern Processing Theory, Defence Research Agency, Malvern WR14 3PS, UK.

Internet: webber@signal.dra.hmg.gb

Abstract

A geometrical interpretation of the elementary constituents which make up perceptual patterns is proposed: if a number of different pattern-vectors lie approximately within the same plane in the pattern-vector space, those patterns can be interpreted as sharing a common constituent. Individual constituents are associated with individual planes of patterns: a pattern lying within an intersection of several such planes corresponds to a combination of several constituents. This interpretation can model patterns as hierarchical combinations of constituents that are themselves combinations of yet more elementary constituents.

A neuron can develop transformation-invariances in its recognition-response by aligning its synaptic vector with one of the plane-normals: a pattern-vector's projection along the synaptic vector is then an invariant of all the patterns on the plane. In this way, discriminating detectors for individual constituents can self-organise through Hebbian adaptation. Transformation-invariances that can self-organise in multiple-level vision systems include shape-tolerance and local position-tolerance.

These principles are illustrated with demonstrations of transformation-tolerant face-recognition.

1. Transformation-invariant neural response

1.1. Geometrical interpretation of invariance

This article interprets invariances of perceptual neurons' pattern-responses in terms of the geometry of the space of pattern-vectors: the orientation of a neuron's synaptic weight-vector with respect to the distribution of pattern-vectors determines how its response varies from pattern to pattern, and so determines for which patterns and under which transformations the response remains the same.

1.2. Pattern-vectors and their frequency-distribution

Neural perception-systems are assemblies of perception-units: neurons which receive their input information as a collection of variable activities at their input synapses. Neurons at the lowest level of a perception-system receive their inputs from an array of sensors such as photoreceptors, while those at higher levels receive input-synapses from other perception-units.

Each neuron receives its input-information encoded in a set of N coordinates

$$\boldsymbol{x} = \left(x_1,\ldots,x_i,\ldots,x_N\right)^{\mathrm{T}} \tag{1}$$

© British Crown Copyright 1994/DRA
Reproduced with the permission of the Controller of Her Britannic Majesty's Stationery Office

which represent the current pattern of the neuron's N input activities: a pattern-vector[1] in an N-dimensional pattern-space. Different activity-patterns arise from the perceptual environment with different probabilities

$$P_{\boldsymbol{x}}(\boldsymbol{x}) \ . \tag{2}$$

The structure of this probability-distribution summarises the statistical information available to the neuron's experience. The neuron adapts to its environment on the basis of its experience of some *ensemble* $\{\boldsymbol{x}\}$ of patterns populated according to this distribution.

1.3. The neuron's response-function and the synaptic vector

A neuron's *pattern-response function* $r(\boldsymbol{x})$ determines the value of the neuron's response to any particular input-activity pattern $\boldsymbol{x}$ that appears at its input-synapses. Real-valued, non-stochastic response-functions are considered in this article. Neural-network models usually assume that $r(\boldsymbol{x})$ depends on the vector[2]

$$\boldsymbol{w} \ = \ \left(w_1,\ldots,w_i,\ldots,w_N\right)^{\mathrm{T}} \tag{3}$$

of the neuron's N input synaptic weights: the synaptic vector, which constitutes the neuron's receptive field profile. The response is usually assumed to be a continuous (non-decreasing) function of the scalar product of $\boldsymbol{x}$ and $\boldsymbol{w}$,

$$r(\boldsymbol{x},\boldsymbol{w}) \ = \ r(\boldsymbol{w}\cdot\boldsymbol{x}) \ . \tag{4}$$

The scalar product $\boldsymbol{w}\cdot\boldsymbol{x}$ depends on the vector-length $|\boldsymbol{w}|$ and on each pattern's *projection*

$$a \ \equiv \ \boldsymbol{x}\cdot\hat{\boldsymbol{w}} \tag{5}$$

along the synaptic vector's unit-direction

$$\hat{\boldsymbol{w}} \ \equiv \ |\boldsymbol{w}|^{-1}\boldsymbol{w} \ . \tag{6}$$

For any given state of the synaptic vector, the pattern-response depends on the pattern-vector only via its projection a:

$$r(\boldsymbol{x})\big|_{\boldsymbol{w}} \ = \ r\big(\boldsymbol{x}\cdot\hat{\boldsymbol{w}}\big) \ . \tag{7}$$

Pattern-vectors having equal projections $a = a^{\mathrm{o}}$ onto $\hat{\boldsymbol{w}}$ lie within the same N–1-dimensional plane

$$\boldsymbol{x}\cdot\hat{\boldsymbol{w}} \ = \ a^{\mathrm{o}} \tag{8}$$

in the pattern space, which is normal to the direction of $\hat{\boldsymbol{w}}$ Pattern-vectors having *similar* projections $a \approx a^{\mathrm{o}}$ onto $\hat{\boldsymbol{w}}$ lie *close* to the plane.

1.4. Transformation-invariance of the response-function

A neuron having the generic response-function (7) requires special orientation of its synaptic vector among the pattern-vectors in order to achieve transformation-invariant pattern-response. Different patterns will induce the *same response* if they have the *same projection* $a = a^{\mathrm{o}}$ along the synaptic vector, *i.e.* if they all lie within the same N–1-dimensional plane normal to $\hat{\boldsymbol{w}}$: the plane defined by (8). Because the response-function is continuous, patterns

---

[1]Pattern-vectors are represented here as column-vectors: $(\cdots)^{\mathrm{T}}$ indicates matrix transposition.

[2]The synaptic vector $\boldsymbol{w}$ is a vector in the N-dimensional pattern-space of the activity-patterns $\boldsymbol{x}$.

that lie close to a plane ($a \approx a^{o}$) will induce closely similar responses. Patterns which have unequal projections onto the synaptic vector will induce different responses.

Pattern-transformations

$$\boldsymbol{x} \xrightarrow{S} \boldsymbol{x}' = \boldsymbol{S}\boldsymbol{x} \tag{9}$$

can describe any relationship between pairs of pattern-vectors ($\boldsymbol{x},\boldsymbol{x}'$). For example, one transformation might change the expression of a face in an image-vector from a smile to a frown, while another might displace a whole face sideways across the image. Of interest here are pattern-transformations that leave the response invariant: they relate different patterns that induce similar responses,

$$r(\boldsymbol{S}\boldsymbol{x})\big|_{w} = r(\boldsymbol{x})\big|_{w} \ . \tag{10}$$

A neuron's pattern-response can be made invariant under a number of transformations if the pattern-vectors related by those transformations lie within a plane: to achieve this, the neuron's synaptic vector must be brought into alignment with the plane's normal.

This is the basis of *discriminating* transformation-invariance: an individual neuron's response can be made invariant under a group of specific pattern-transformations, whilst discriminating between patterns that are related to one another by other transformations. Self-organisation of discriminating transformation-invariance relies on adaptive alignment of synaptic vectors, in such a way that they become normal to planes of pattern-vectors experienced.

2. Shared constituents of patterns

2.1. Interpreting the building-blocks of an ensemble of patterns

The approach to self-organised invariance presented here is relevant for ensembles of patterns that are composed from combinations of elementary *constituents*: sub-structures that recur in various combinations as the shared building-blocks of the ensemble's various patterns. This article proposes a geometrical definition to interpret the intuitive concept of pattern-constituents and clarify their significance for invariant response, and presents a neural adaptation mechanism which generates detectors for them. Within this framework, a pattern-ensemble's constituent-structure is the statistical property which permits individual neurons to respond with transformation-invariance.

2.2. Pattern-constituents: definitions

Various pattern-subsets within an ensemble may be coplanar to some approximation. A *constituent* is a shared sub-structure associated with each pattern of such a coplanar subset.

Suppose some subset of pattern-vectors $\boldsymbol{x}$ drawn from the ensemble $\{\boldsymbol{x}\}$ lie on the same N–1-dimensional plane

$$\boldsymbol{x}\cdot\hat{\boldsymbol{c}}_i = a_i^{o} \ . \tag{11}$$

Associated with each such plane i of pattern-vectors is a vector $\hat{\boldsymbol{c}}_i$ normal to the plane. A constituent will be defined as a (unit) vector[3] $\hat{\boldsymbol{c}}_i$ normal to a plane passing through some such subset of the pattern-ensemble. The definition is readily extended to subsets of patterns that

[3]The ensemble's constituents are vectors in the pattern-space, though they are not generally pattern-vectors belonging to the ensemble $\{\boldsymbol{x}\}$.

are coplanar only within a finite approximation. The projection of any given pattern $\boldsymbol{x}$ along a constituent $\hat{\boldsymbol{c}}_i$ is defined as

$$a_i \equiv \boldsymbol{x} \cdot \hat{\boldsymbol{c}}_i \ . \tag{12}$$

(This is analogous to extracting Fourier components, which are projections along unit basis-vectors.) Also associated with each plane is its defining projection a_i^{o}: all the patterns lying on a constituent's normal-plane $\boldsymbol{x}\cdot\hat{\boldsymbol{c}}_i = a_i^{\mathrm{o}}$ cast *similar* projections

$$a_i \ = \ a_i^{\mathrm{o}} \tag{13}$$

along the constituent. This discussion is only appropriate for pattern-distributions $P_{\boldsymbol{x}}(\boldsymbol{x})$ having dense parts which lie on or near planes, though the examples below demonstrate that such non-random structure arises in realistic situations.

A pattern-distribution $P_{\boldsymbol{x}}(\boldsymbol{x})$ characterised by constituents will cast projections $\boldsymbol{x}\cdot\hat{\boldsymbol{c}}_i$ which follow distributions $P_{a_i}(a_i)$ peaked at a_i^{o}: the peak contains the patterns on the constituent's normal-plane, while the other patterns of the ensemble form a background at other a_i values. The coplanar patterns in the peak can be selected from most or all of the rest of the ensemble by demanding a threshold-cut in a_i, at some value ϑ which is less[4] than a_i^{o}:

$$\boldsymbol{x} \cdot \hat{\boldsymbol{c}}_i \ > \ \vartheta \qquad\qquad (\vartheta < a_i^{\mathrm{o}}) \ . \tag{14}$$

The cut just divides the pattern-space along the plane $\boldsymbol{x}\cdot\hat{\boldsymbol{c}}_i = \vartheta$, which is parallel to the coplanar pattern-subset and separated from it by a distance $a_i^{\mathrm{o}} - \vartheta$. The value of ϑ could be chosen as close to a_i^{o} as the coplanarity's accuracy permits. (Approximate coplanarity arises when pattern-vectors lie on a smoothly-curved convex surface for example: patterns lying near the surface's tangent-plane at $\hat{\boldsymbol{c}}_i$ will be approximately coplanar, and can be separated from the other patterns by a planar threshold-cut.)

A further piece of terminology is useful shorthand for discussing pattern-constituents, in situations where no pattern in the ensemble casts a projection a_i larger than a_i^{o}, and it is only significant to specify whether or not a given pattern lies in the coplanar subset $\boldsymbol{x}\cdot\hat{\boldsymbol{c}}_i = a_i^{\mathrm{o}}$. With respect to some threshold-plane $\boldsymbol{x}\cdot\hat{\boldsymbol{c}}_i = \vartheta$ chosen as above, a pattern $\boldsymbol{x}$ is said to *contain* constituent i, or equivalently constituent i is *present* within pattern $\boldsymbol{x}$, if $\boldsymbol{x}$ satisfies (14). Conversely, constituent i is *absent* from pattern $\boldsymbol{x}$ if $\boldsymbol{x}$ lies on the other side of this threshold-plane. (The threshold-cut (14) separates the patterns containing $\hat{\boldsymbol{c}}_i$ from the patterns in which $\hat{\boldsymbol{c}}_i$ is absent.)

By this definition, all the patterns in the coplanar subset $\boldsymbol{x}\cdot\hat{\boldsymbol{c}}_i = a_i^{\mathrm{o}}$ defining constituent $\hat{\boldsymbol{c}}_i$ will contain that constituent: the constituent is *shared* by all those different patterns, recurring in pattern after pattern. If any particular pattern happens to be a member of several distinct (non-parallel, intersecting) coplanar subsets of the ensemble, that pattern will contain several distinct constituents.

---

[4]The sign convention for $\hat{\boldsymbol{c}}_i$'s direction can always be defined so that this cut separates the a_i^{o} peak from the bulk of the ensemble's patterns.

2.3. Invariant representation of an ensemble's patterns by their constituents

A pattern-ensemble that contains many distinct coplanar subsets has many distinct constituents. The kind of high-dimensional pattern ensemble of interest here has large numbers of constituents, several of which are present in most or all of the ensemble's patterns: different *combinations* of constituents are present in different individual patterns. The complete set $\{\hat{\boldsymbol{c}}_i\}$ of an ensemble's constituents forms a vector-basis along which a pattern-vector's projections $\{a_i\}$ make up a transformed set of coordinates for representing the pattern. The number of distinct constituents is not constrained to be equal to, less than, or greater than N, so the transform-basis is not required to be complete or linearly independent. Permitting the basis to be non-orthogonal (as in Gabor-wavelet representation, for example [*e.g.* Field 1993]) allows it to consist of subtle, information-rich constituents: because the transform-basis is free to be non-orthogonal and the transform-coordinates non-linear, the transform can be much more general than mere pattern-space rotation.

The coplanar patterns which all contain a particular constituent all cast similar projections along that constituent. In the transformed representation, patterns sharing a common constituent $\hat{\boldsymbol{c}}_i$ can be recognised from the corresponding coordinate a_i, which will be invariant for all those patterns: an invariant under the transformations that relate different patterns containing the particular constituent. The constituent-representation is a transformation-invariant representation: one in which each transform-coordinate is an invariant under a distinct group of transformations mapping ensemble-patterns onto one another.

A pattern's constituent-representation is a list specifying which combination of constituents is present in the pattern. In information language [*e.g.* Shannon 1949], the complete set of the ensemble's constituents corresponds to an `alphabet of symbols' from which the patterns of the ensemble are built up in various combinations.

2.4. Illustration: a visual pattern-ensemble

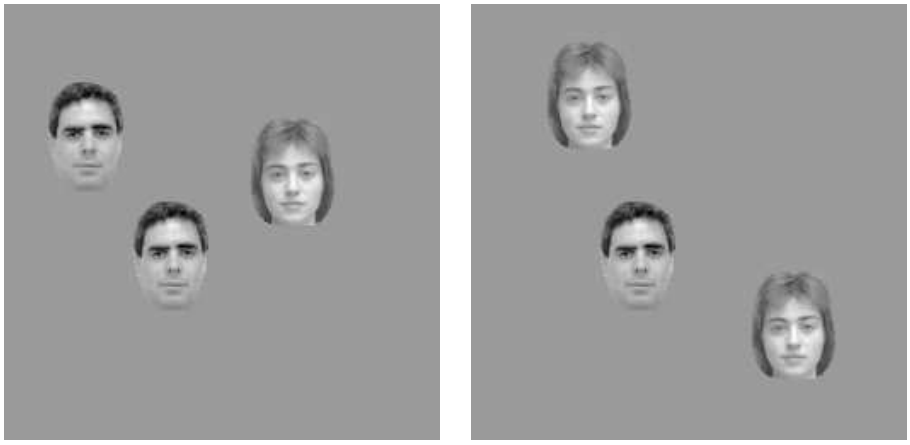

Figure 1: Examples from an ensemble of patterns assembled from random combinations of a few elementary constituents, selected from an alphabet of available constituents. Each constituent of the alphabet is one of two people's faces, at a specific image-location. (There are two constituents of the alphabet for every image-location.) The two patterns depicted here happen to share a single constituent in common (one face at a particular location): other pairs and sets of patterns in the ensemble share other constituents in common.

Consider an illustration: a visual pattern-ensemble in which each distinct pattern-vector represents a different 256×256 grey-level image. Each constituent is a recognisable sub-structure which recurs within pattern after pattern: an individual face, of a particular person,

at a specific image-location. The complete alphabet consists of two different faces (two people), each at every possible image-location (2×256×256 distinct constituents in total). From combinations of constituents selected from this alphabet can be assembled huge numbers of distinct photo-montages of groups of faces, such as the two depicted in figure 1. Each pattern of the ensemble is a different combination of a few faces selected from the alphabet (satisfying the constraint that no *single* pattern contains faces that overlap with one another). Each individual constituent is an image-space vector: an image containing just one face (figure 3 depicts two of these constituent-vectors).

Each constituent is an axis along which all the ensemble's patterns containing the constituent cast similar projections. Notice that the two patterns of figure 1, which both contain the first constituent depicted in figure 3, cast similar projections along this axis: so do all the other *coplanar* patterns that contain this constituent. All the other constituents of the alphabet define other planes of patterns, which share other faces in common. A pattern containing a particular combination of faces lies on the intersection of the planes defined by those constituents.

The alphabet of constituents is the basis of a discriminating, transformation-invariant representation of the ensemble's patterns. Each coordinate discriminates between patterns exclusively on the basis of whether they contain a *particular* face, and is an invariant with respect to transformations of other faces.

2.5. Intersection/Union relationship between patterns and constituents

Regarding patterns as sets, within the universal set of the whole ensemble $\{\boldsymbol{x}\}$, illuminates the relationship between patterns and constituents. Any subset of the ensembles' patterns lying in a plane defines a constituent, which all those patterns contain. This subset, associated with an individual constituent, is the union $\bigcup$ of all the patterns containing the constituent. Conversely, an individual pattern is the intersection $\bigcap$ of all the pattern-subsets associated with constituents the pattern contains. In the pattern-space, the pattern-vector which contains a particular combination of constituents lies on the *geometrical* intersection between the corresponding combination of constituent-planes. Patterns are particular combinations of constituents in that they are particular intersections of constituent-defining sets.

2.6. Composite constituents: intersections of elementary constituents

In the pattern-set picture, an individual pattern is a constituent-intersection that contains just one member. Pattern-subsets having more than one member can also be intersections of constituent-defining subsets. For example, the subset of the ensemble's patterns containing constituent $\hat{\boldsymbol{c}}_1$ includes the *smaller* subset of patterns that contain both $\hat{\boldsymbol{c}}_1$ and $\hat{\boldsymbol{c}}_2$ in conjunction, which in turn includes the subset containing $\hat{\boldsymbol{c}}_1$, $\hat{\boldsymbol{c}}_2$ and $\hat{\boldsymbol{c}}_3$. The subset containing $\hat{\boldsymbol{c}}_1$ and $\hat{\boldsymbol{c}}_2$ is the intersection between the subset containing $\hat{\boldsymbol{c}}_1$ and the subset containing $\hat{\boldsymbol{c}}_2$. Similarly, the patterns containing the conjunction $\hat{\boldsymbol{c}}_1,\hat{\boldsymbol{c}}_2,\hat{\boldsymbol{c}}_3,\ldots$ make up the *multiple* intersection of the subsets that each contain $\hat{\boldsymbol{c}}_1,\hat{\boldsymbol{c}}_2,\hat{\boldsymbol{c}}_3,\ldots$ individually.

The subset of patterns that contain a given constituent $\hat{\boldsymbol{c}}_i$ with projection a_i^{o} all lie on the plane $\boldsymbol{x}\cdot\hat{\boldsymbol{c}}_i = a_i^{\mathrm{o}}$. Consequently, all the patterns containing the conjunction $\hat{\boldsymbol{c}}_1,\hat{\boldsymbol{c}}_2,\hat{\boldsymbol{c}}_3,\ldots$ lie on the intersection of the planes defined by $\boldsymbol{x}\cdot\hat{\boldsymbol{c}}_1 = a_1^{\mathrm{o}}$, $\boldsymbol{x}\cdot\hat{\boldsymbol{c}}_2 = a_2^{\mathrm{o}}$, $\boldsymbol{x}\cdot\hat{\boldsymbol{c}}_3 = a_3^{\mathrm{o}},\ldots$ (which exists provided the vectors $\hat{\boldsymbol{c}}_1,\hat{\boldsymbol{c}}_2,\hat{\boldsymbol{c}}_3,\ldots$ form a linearly independent set). The patterns within this intersection of planes must be mutually coplanar, and so they define a new constituent, which is distinct from all of $\hat{\boldsymbol{c}}_1,\hat{\boldsymbol{c}}_2,\hat{\boldsymbol{c}}_3,\ldots$ [Appendix A.]. The patterns containing this new constituent all contain the conjunction $\hat{\boldsymbol{c}}_1,\hat{\boldsymbol{c}}_2,\hat{\boldsymbol{c}}_3,\ldots$: this new constituent is a *composite* of more

elementary sub-constituents $\hat{\boldsymbol{c}}_1,\hat{\boldsymbol{c}}_2,\hat{\boldsymbol{c}}_3,\ldots$. A composite constituent will be present more *rarely* than will its elementary sub-constituents, because the intersection of subsets contains fewer of the ensemble's patterns than does each elementary subset. The patterns of an intersection must always cast a projection along the composite constituent which is larger than any of the projections $a_1^o,a_2^o,a_3^o,\ldots$ along the elementary constituents [Appendix A.].

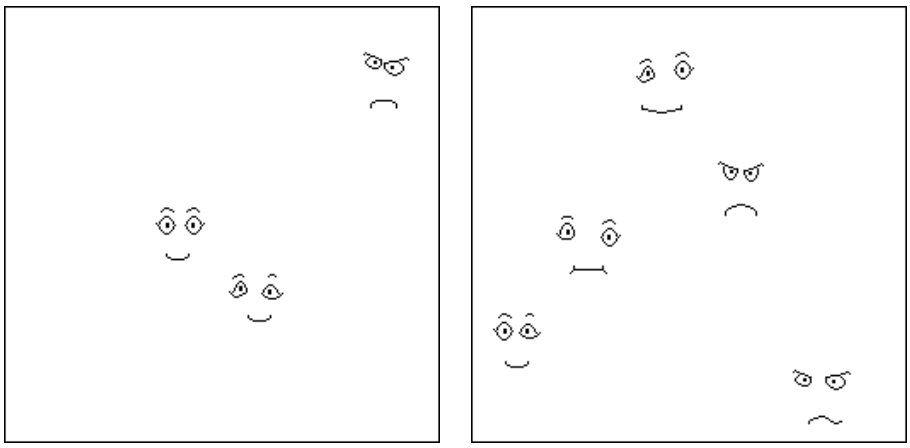

Figure 2: Examples from another ensemble of patterns, assembled from random combinations of faces that vary in expression and shape: each face is constructed as a variable combination of mouths and left and right eyes with *various alternative expressions*, and with *variable relative separations*. The faces wearing all these many expressions each belong to one of two cartoon-'people': one person whose face usually contains smiling eyes and mouths (with varying degrees of smile), and a second person whose face usually contains frowning eyes and mouths (with varying degrees of frown). Each individual eye or mouth that recurs in several patterns is a shared constituent of those patterns. Patterns of the ensemble are combinations of *composite* constituents (faces with specific expressions), that are themselves combinations of elementary sub-constituents (specific eyes and mouths).

The concept of composite constituents requires an illustration: an extension of the face-montage pattern-ensemble [Figure 2]. In the images of this new ensemble, the faces of the two people are allowed to vary in shape and expression, as well as in position as before. Different shape-variants of eyes and mouth make up faces with different expressions, while differently-shaped faces are composed from eyes and mouths with different relative displacements. Each individual composite face is a *particular* combination of 3 elementary facial features, and there are many such distinct composites, each one arising in patterns much more rarely than does any particular elementary facial feature. (See ahead to figures 5(*a*) and 5(*b*) for illustrations of these composite and elementary constituents.) Variations in the *combinations* of elementary constituents that make up the composites are responsible for variation in face-shape and expression.

Composite constituents can have hierarchical structures of many levels: a composite may be the conjunction of several more elementary constituents, which may in turn be composites of yet more elementary constituents. Identifying pattern-constituents with coplanar subsets of the pattern-ensemble enables us to model the intuitive notion that the perceptual world is made up of hierarchies of features within features, and to interpret the transformation-invariances involved in recognising those features.

2.7. Constituent-representation as an unredundant sparse code

For ensembles of patterns concentrated on constituent-planes, the constituent-representation $\{a_i\}$ is a sparse representation of patterns: each pattern's a_i-coordinates will be large (greater than some threshold ϑ) for only a few constituents i out of the alphabet. The

representation is far less redundant [Barlow 1959] than the input-coordinate representation $\{x_i\}$, because only a small proportion of the alphabet's symbols are required to describe any individual pattern, on average.

Real perceptual environments seem suitable for sparse, unredundant representation in terms of frequently-arising sub-structures. Neurons of the visual cortex may compute (functions of) the coordinates of a sparse transform of natural scenes [Field 1987]. Using a unredundant neural code for representing frequent sub-structures is likely to be a primary function of vertebrates' brains [Barlow 1972]: tuning individual neurons to frequent sub-structures would be achieved through adaptation.

3. Synaptic adaptation through invariance self-organisation

3.1. The goal of invariance self-organisation

If neurons having the generic response-function (7) are to transform their input-patterns into an invariant constituent-representation $\{a_i\}$, their synaptic vectors should adapt so as to become aligned with the constituents of the pattern-ensemble they experience. A neuron having synaptic vector aligned with constituent $\hat{\boldsymbol{c}}_i$ will have a response-function determined by the input-pattern's corresponding projection a_i:

$$r(\boldsymbol{x})\big|_{\boldsymbol{w}\|\hat{\boldsymbol{c}}_i} = r(a_i) \ . \tag{15}$$

Having aligned its synaptic vector with constituent $\hat{\boldsymbol{c}}_i$, a neuron's response will be invariant for the coplanar subset of patterns which contain $\hat{\boldsymbol{c}}_i$.

An adaptive dynamics called *invariance self-organisation* (ISO) drives synaptic vectors into alignment with constituents. ISO was designed specifically to exhibit this behaviour, which can be attributed to the form of its dynamical equation. That this dynamical equation implicitly contains a simple neuron-model is coincidental to this primary goal.

3.2. The response-function of the ISO neuron

The ISO dynamical equation contains an implicit model for the neural response-function $r(\boldsymbol{x})\big|_{\boldsymbol{w}}$. The response to an activity-pattern $\boldsymbol{x}$, for a given state $\boldsymbol{w}$ of the input-synapses, is determined as

$$r(\boldsymbol{x})\big|_{\boldsymbol{w}} = \theta(\boldsymbol{x}\cdot\boldsymbol{w} - \vartheta|\boldsymbol{w}|) = |\boldsymbol{w}|\theta(\boldsymbol{x}\cdot\hat{\boldsymbol{w}} - \vartheta) \ , \tag{16}$$

where a non-linear (piecewise linear) threshold-function $\theta(\zeta)$ is defined

$$\theta(\zeta) \equiv \begin{cases} \zeta \text{ if } \zeta \geq 0 \\ 0 \text{ if } \zeta \leq 0 \end{cases} \ . \tag{17}$$

The ISO response-function bears a superficial resemblance to the classical Perceptron activation-function [Minsky & Papert 1969] (though their learning-rules are unrelated).

The (scalar) *threshold-parameter* ϑ is a fixed property of each individual neuron, which does not change[5] during the self-organisation of the synaptic vector. It has the dimensions of a length in pattern-space. This response-function belongs to the generic class (7): the response depends on the pattern only via its projection along $\hat{\boldsymbol{w}}$.

---

[5] ϑ is not a dynamical variable of ISO. However, ϑ may be allowed to change very gradually in order to follow the paths of ISO convergent states [Section 3.7.].

3.3. Selective response: the ISO threshold-parameter ϑ

A neuron with this response-function will respond only to patterns that cast projections greater than ϑ onto $\hat{\boldsymbol{w}}$:

$$r(\boldsymbol{x})\big|_{\boldsymbol{w}} > 0 \qquad \Leftrightarrow \qquad \boldsymbol{x}\cdot\hat{\boldsymbol{w}} > \vartheta \ . \tag{18}$$

The response-function defines the neuron's *response-sector*: a region of the pattern space which lies to one side of the threshold-plane $\boldsymbol{x}\cdot\hat{\boldsymbol{w}} = \vartheta$. Only patterns that fall within this sector will induce a response (18), which will be proportional to the pattern's distance from the threshold-plane. (The subset of the ensemble's patterns that fall within this sector will be denoted by $\Theta(\hat{\boldsymbol{w}},\vartheta)$: this response-inducing subset depends on $\hat{\boldsymbol{w}}$ and ϑ according to (18).)

The distance ϑ of this threshold-plane from the origin determines the neuron's degree of pattern-selectivity (how often it responds), because a more distant response-sector will contain a smaller subset $\Theta(\hat{\boldsymbol{w}},\vartheta)$ of the ensemble's patterns. It follows from (18) that the range of useful ϑ values is restricted to

$$-\max_{\{\boldsymbol{x}\}}|\boldsymbol{x}| < \vartheta < +\max_{\{\boldsymbol{x}\}}|\boldsymbol{x}| \ . \tag{19}$$

With ϑ below the lower limit, the neuron will respond to any pattern in the ensemble $\{\boldsymbol{x}\}$: if ϑ is above the upper limit, the neuron will never respond. The response becomes progressively more pattern-selective as the threshold-parameter ϑ is increased.

If the synaptic vector's direction $\hat{\boldsymbol{w}}$ becomes aligned with one of the ensemble's constituents $\hat{\boldsymbol{c}}_i$, and if the ISO threshold satisfies

$$\vartheta < a_i^{\mathrm{o}} \ , \tag{20}$$

the response-sector will contain all the ensemble's coplanar patterns that contain $\hat{\boldsymbol{c}}_i$: the threshold-plane then lies parallel to that particular constituent-plane. The neuron's response will then be selective for patterns containing that constituent, and invariant for those patterns: the neuron will be a discriminating, invariant detector for a specific constituent.

3.4. The dynamical equation of invariance self-organisation

ISO prescribes adaptation of the synaptic vector to an ensemble of patterns $\{\boldsymbol{x}\}$ according to dynamics

$$\frac{d\boldsymbol{w}}{dt} = \lambda\left\langle \boldsymbol{x}\,\theta(\boldsymbol{w}\cdot\boldsymbol{x} - \vartheta|\boldsymbol{w}|)\right\rangle_{\{\boldsymbol{x}\}} \qquad (\text{adaptation rate } \lambda > 0) \ . \tag{21}$$

The average is over the frequency-distribution $P_{\{\boldsymbol{x}\}}(\boldsymbol{x})$ of patterns in the ensemble. The ISO dynamics (21) perform gradient ascent [Appendix C.] on a potential $V(\boldsymbol{w})$, which is the mean squared distance of the above-threshold patterns $\Theta(\hat{\boldsymbol{w}})$ from the threshold-plane $\hat{\boldsymbol{w}}\cdot\boldsymbol{x} = \vartheta$. (These dynamics can be simulated by iteration at discrete time-steps $dt = 1$, provided the adaptation rate λ is sufficiently small that, at each step, $\boldsymbol{w}$ is incremented by a vector $d\boldsymbol{w}$ small in comparison with $\boldsymbol{w}$: this condition is met when $1/\lambda$ is substantially larger than the variance of the ensemble's pattern-vectors. In this small-λ limit, the trajectory followed becomes independent of λ, as with gradient-following on any continuous function: all the numerical simulations presented in this article were made in the λ-independent limit.)

The dynamical equation (21) is homogeneous: it is left invariant in form by any operation that rescales the synaptic vector

$$\boldsymbol{w} \ \rightarrow \ \boldsymbol{w}' = s\boldsymbol{w} \qquad (\text{scalar } s) \ . \tag{22}$$

In other words, the adaptation of $\boldsymbol{w}$'s *direction* is unaffected by any process which changes only $\boldsymbol{w}$'s *length*. Consequently, $\boldsymbol{w}$ may be rescaled continuously throughout its adaptation, without interfering with the dynamics, in order to conserve its normalisation at any desired value. Homogeneity therefore allows a normalising mechanism to operate independently of the basic adaptation process: whatever normalisation scheme operates, the unit vector $\hat{\boldsymbol{w}}$ follows [Appendix B.]

$$\frac{d\hat{\boldsymbol{w}}}{dt} = \lambda \left\langle [\boldsymbol{x} - (\hat{\boldsymbol{w}} \cdot \boldsymbol{x})\hat{\boldsymbol{w}}]\theta(\hat{\boldsymbol{w}} \cdot \boldsymbol{x} - \vartheta) \right\rangle_{\{\boldsymbol{x}\}} . \tag{23}$$

This property is familiar from the linear principal component (PCA) neuron [Oja 1982], which has scale-invariant dynamics: the ISO dynamics differ in having the scale-preserving but strongly non-linear function $\theta(\zeta)$.

The ISO dynamics are also invariant with respect to the units in which *pattern-vectors*' lengths are measured, as long as the ϑ-parameter is consistently measured in those same length-units. (The dynamical equation (21) is invariant under rescaling of $\boldsymbol{x}$ by the same factor by which ϑ is rescaled.) This means that the convergent states do not depend on the normalisation of the data.

These dynamics are equivalent to Hebbian adaptation of the ISO neuron [Hebb 1949]. In response to pattern $\boldsymbol{x}$, drawn from $P_{\{\boldsymbol{x}\}}(\boldsymbol{x})$ and persisting for duration δt, the synaptic vector is augmented towards $\boldsymbol{x}$ by an amount proportional to the neuron's response:

$$\delta \boldsymbol{w}(\boldsymbol{x}) = \lambda \delta t \boldsymbol{x}\, r(\boldsymbol{x})\big|_{\boldsymbol{w}} . \tag{24}$$

In the small-δt limit, in which the sampled mean $\delta t \sum_{n=1}^{1/\delta t} \boldsymbol{x}_n\, r(\boldsymbol{x}_n)$ asymptotes to the true mean $\langle \boldsymbol{x}\, r(\boldsymbol{x}) \rangle_{\{\boldsymbol{x}\}}$, the stochastic recursion (24) will follow the path of the deterministic dynamics (21), converging to its stable stationary points. (Under these circumstances, general conditions [Ljung 1977] for the stochastic recursion to converge to the stable points of the deterministic dynamics are satisfied.)

Like the linear PCA neuron, but unlike other unsupervised structure-finding networks such as competitive nets [Rumelhart & Zipser 1985], learning vector-quantisers or topographic feature-maps [Kohonen 1989], individual ISO neurons self-organise independently: without (lateral) interaction with one another. Consequently, a set of ISO neurons adapting to the same pattern-vectors will have no ordering or topographic neighbourhood-relationships. ISO neurons in multiple-level stacks are oblivious to this lack of ordering, since no special ordering is assumed between the components of any neuron's input-vector.

3.5. ISO convergence to organised states

Provided that $\hat{\boldsymbol{w}}$ is initially oriented so that at least one of the ensemble's patterns lies above threshold (18) at the beginning of self-organisation, these dynamics inevitably bring the synaptic vector's direction to convergence

$$\lim_{t\to\infty} \hat{\boldsymbol{w}}(t) = \hat{\boldsymbol{w}}^{\infty} \qquad \lim_{t\to\infty} \frac{d\hat{\boldsymbol{w}}(t)}{dt} = \boldsymbol{0} , \tag{25}$$

oriented after convergence so that some patterns of the ensemble will induce above-threshold response [Appendix D.].

For a given pattern-distribution, many alternative $\hat{\boldsymbol{w}}^{\infty}$s are convergent states of the ISO dynamics, though any single neuron cannot converge to more than one. Which alternative is actually selected depends on which basin of attraction happened to contain the initial $\hat{\boldsymbol{w}}$ at the outset of adaptation. A population of neurons, adapting independently of one

another, can converge to different states from the available set $\{\hat{\boldsymbol{w}}^{\infty}\}$: all these convergent states will be represented in the population if it is large enough, and if the initial states were distributed sufficiently uniformly among the various basins of attraction at the outset.

3.6. ISO converges on constituents selected by ϑ

An ensemble's constituents $\{\hat{\boldsymbol{c}}_i\}$ are convergent vectors $\{\hat{\boldsymbol{w}}^{\infty}\}$ of ISO adaptation to the ensemble. The general conditions [Appendix E.] for $\hat{\boldsymbol{w}}^{\infty}$ to be a stable convergent state of ISO take a simple form [Appendix F.] if $\hat{\boldsymbol{w}}^{\infty}$ is aligned with a constituent $\hat{\boldsymbol{c}}_i$: $\hat{\boldsymbol{c}}_i$ will correspond to a convergent state if the coplanar patterns $\Theta: \boldsymbol{x}\cdot\hat{\boldsymbol{c}}_i = a_i^{\mathrm{o}}$ which define the constituent have their mean lying along $\hat{\boldsymbol{c}}_i$

$$\langle \boldsymbol{x} \rangle_{\Theta} \;=\; a_i^{\mathrm{o}} \hat{\boldsymbol{c}}_i \;, \tag{26}$$

and if their principal variance v_{Θ}^{1} (the principal eigenvalue of their covariance matrix) satisfies a stability-condition

$$v_{\Theta}^{1} \;<\; a_i^{\mathrm{o}}\left(a_i^{\mathrm{o}} - \vartheta\right), \qquad i.e. \qquad \vartheta \;<\; a_i^{\mathrm{o}} - v_{\Theta}^{1} \big/ a_i^{\mathrm{o}} \;. \tag{27}$$

A pattern-subset need only be approximately coplanar to give rise to such a convergent state: if a subset Θ of patterns lie *close* to the plane $\boldsymbol{x}\cdot\hat{\boldsymbol{c}}_i = a_i^{\mathrm{o}}$, a convergent state $\hat{\boldsymbol{w}}^{\infty}$ will lie close to $\hat{\boldsymbol{c}}_i$ [Appendix G.].

The stability-condition (27) sets an upper bound on the ISO parameter ϑ, above which the convergent state bifurcates[6] [Appendix F.]. If v_{Θ}^{1} is small[7] compared to $a_i^{\mathrm{o}2}$, the stationary point $\hat{\boldsymbol{w}}^{\infty}$ aligned with $\hat{\boldsymbol{c}}_i$ will be stable unless ϑ is raised almost as high as a_i^{o}. A constituent will be a convergent state within a range of ϑ, bounded above by (27), and bounded below by the requirement that the response-sector excludes patterns outside Θ. Because a composite constituent has a defining projection a_i^{o} larger than the defining projections associated with the elementary constituents comprising it [Appendix A.], it will be a stable convergent state at higher ϑ values than those at which its elementary constituents are stable.

As ϑ is increased through its full range (19), bifurcation-transitions between several distinct phases will typically be passed, within which constituents with different a_i^{o} are stable. ϑ serves as a control, to select between different constituents associated with different defining projections a_i^{o}.

3.7. Path-following of convergent states from low ϑ to high ϑ

ISO convergence is inevitable from any initial state $\hat{\boldsymbol{w}}^{\mathrm{o}}$, provided that $\hat{\boldsymbol{w}}^{\mathrm{o}}$ is not stranded without patterns in its response-sector [Section 3.5.]. At high ϑ-values, response-sectors can contain many fewer patterns than at low ϑ, so it can be much less easy to avoid stranding at high ϑ than at low ϑ if the initial state is chosen randomly. This potential problem is avoided by choosing high-ϑ initial states non-randomly: by following the paths of convergent states through their bifurcations, as ϑ is raised from a low value (where an unstranded initial state can be chosen randomly). Path-following can be employed during

---

[6] See [*e.g.* Stewart 1992] for a general description of bifurcation of differential equations' fixed points.

[7] In spaces of high dimension N, the total variance of an ensemble can be equipartitioned among as many as N mutually orthogonal projections, so principal variances can be very small.

self-organisation if ϑ is raised gradually, over a time-scale substantially longer than the convergence time-scale, so that the state is never perturbed far from convergence.

4. Demonstration of ISO: multiple-level visual learning

4.1. A pattern-distribution for demonstrating ISO in a single level of neurons

The face-constituent pattern-ensemble that was used to illustrate the coplanarity principle [Section 2.4., Figure 1] can be used to demonstrate invariance self-organisation. This pattern-distribution serves as a model of real perceptual information, whilst having statistics sufficiently clearly defined to allow ISO's principles to be visualised geometrically.

The individual patterns consist of combinations of several constituents (a mean of 3 per image) selected randomly from an alphabet of $2\times256\times256$: the faces of 2 equiprobable people, distributed uniformly across the 256×256 image field[8]. Each constituent of the alphabet overlaps many others to a greater or lesser extent, though no two overlapping faces are permitted to be present within the same image. The faces are selected with statistical near-independence: consequently, the number n of faces per image closely follows a Poisson distribution with $\overline{n} = 3$. The images have a uniform background, which is chosen to be zero. Whenever any particular face is present in a pattern, the pattern casts a projection a^{o} along that constituent-axis: a^{o} defines the length-scale of the pattern-space (a pattern containing n faces has vector-length $\sqrt{n}\,a^{o}$, since its n constituents are orthogonal if the background is zero).

4.2. Development of discriminating, transformation-invariant constituent-detectors

Figure 3: Two of the $2\times256\times256$ alternative convergent synaptic vectors that can be obtained from ISO adaptation to a large ensemble of images of *combinations* of faces at various image-locations (like the images of figure 1), with threshold parameter $\vartheta = 0.95\,a^{o}$. Numerically simulated self-organisation, initiated from two different, random unadapted states. Each of these different synaptic vectors is the receptive-field profile of a neuron that responds to the presence of a *specific* face at a *specific* location. At this ϑ-value, self-organisation generates transformation-invariant detectors for *individual constituents* that recur in the pattern-ensemble: a detector responds invariantly to all those different patterns that contain the key constituent which matches its receptive field.

[8]The image field has the usual toroidal periodic boundary condition to avoid the complication of borders.

When the ISO parameter ϑ is set to $0.95a^{\mathrm{o}}$, a neuron's response-sector will exclude all patterns of the face-montage ensemble except those having projections as large as a^{o} along $\hat{\boldsymbol{w}}$. A pattern can only cast such a large projection if $\hat{\boldsymbol{w}}$ is aligned with a face that the pattern contains. Path-following [Section 3.7.] of convergent states to $\vartheta = 0.95a^{\mathrm{o}}$ will therefore bring any neuron's synaptic vector into alignment with one of the 2×256×256 faces of the constituent-alphabet: at $\vartheta = 0.95a^{\mathrm{o}}$, there are 2×256×256 alternative convergent states available for $\hat{\boldsymbol{w}}$. Figure 3 shows two of them, obtained by numerical simulation of the ISO dynamics (21), with path-following from two different low-ϑ random initial states. All 2×256×256 constituents of the alphabet are equally likely outcomes of ISO adaptation, if the random initial states are distributed uniformly among the 2×256×256 basins of attraction, which have equal volumes.

Each of these organised neurons is so highly selective that it responds only to patterns that contain (among others) the constituent which matches its synaptic vector exactly. Notice that all the patterns which can induce a response (those which contain the right face at the right place) have equal projections a^{o} onto the organised synaptic vector. Consequently, the neuron's response is *invariant* under the transformations which relate the different patterns containing the trigger-constituent, and *zero* for patterns not containing the trigger-constituent: it discriminates between patterns purely on the basis of its trigger-constituent's presence or absence.

4.3. Detectors for ranges of similar constituents: Generalisation

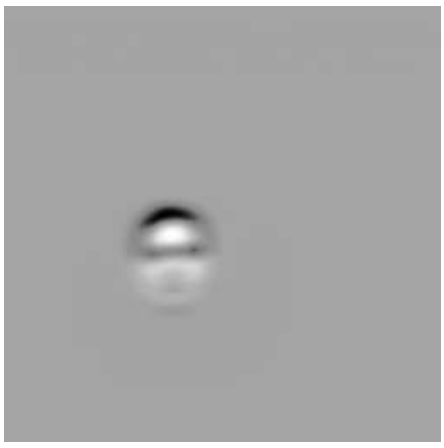

Figure 4: With the threshold relaxed to $\vartheta = 0.5a^{\mathrm{o}}$, the organised ISO neuron responds to patterns containing *any one of a range of similar constituents*: a range of overlapping faces can each induce a response. The convergent synaptic vector is a smooth interpolation between the constituents of that range (an average weighted by response). The response generalises over that range of constituents. (Numerical simulation with path-following from low ϑ.)

Each face of the alphabet overlaps with others to varying degrees: each constituent $\hat{\boldsymbol{c}}_i$ casts projections $\hat{\boldsymbol{c}}_i \cdot \hat{\boldsymbol{c}}_j$ of varying lengths onto other constituents $\hat{\boldsymbol{c}}_j$, because some pairs are more nearly parallel ($\hat{\boldsymbol{c}}_i \cdot \hat{\boldsymbol{c}}_j \approx 1$) than others. Patterns that lie *on* one constituent-plane $\boldsymbol{x} \cdot \hat{\boldsymbol{c}}_i = a_i^{\mathrm{o}}$ will lie *close* to any other constituent-plane $\boldsymbol{x} \cdot \hat{\boldsymbol{c}}_j = a_j^{\mathrm{o}}$ that is nearly parallel to the first: the collection of patterns that contain *any* constituent out of a set of nearly parallel alternatives will all lie approximately in the same plane. Consequently, the response-sector of a detector for patterns on one of these specific planes can be extended, by reducing ϑ, so as to include all these other, less-coplanar patterns as well: as ϑ is progressively reduced, the constituent-

detector will begin to respond to constituents from a *range* of nearly parallel alternatives, which will become broader as ϑ is further reduced.

Figure 4 shows one of the $2\times256\times256$ convergent states at $\vartheta = 0.5a^{\circ}$. Whereas with ϑ only just below a° the neuron responded to just one specific face, it now responds to faces from a range of alternatives, which are similar (nearly parallel). The convergent synaptic vector $\hat{\boldsymbol{w}}^{\infty}$ is a smooth interpolation between these similar constituents: they cast (approximately) equal projections onto $\hat{\boldsymbol{w}}^{\infty}$. Consequently, the response remains approximately invariant whichever of these alternative constituents happens to be present in the pattern: the response *generalises* over this range of similar constituents. This is because the patterns which contain *any* of these alternatives form an approximately coplanar subset: the smooth interpolation is a constituent-vector in its own right.

4.4. A composite pattern-distribution for demonstrating multiple-level ISO

In order to represent with discriminating invariance a pattern-ensemble having *composite* constituents, more than one level of ISO neurons is required. Each level represents constituents that are composites of the constituents encoded at the level below, generalising over the variability in their composition.

This can be demonstrated using the ensemble of montages of variable cartoon-faces that illustrated the principle of composite constituents [Section 2.6., Figure 2]. Cartoon-faces rather than real photographs are used in this demonstration, because variable combinations of cartoon eyes and mouths can be generated in large numbers according to well-defined ensemble statistics. The faces of two cartoon-people, denoted S (smiling) and F (frowning), are distributed with the same near-independent statistics as the faces of two people in the photo-montage ensemble [Section 4.1.]. These faces are each composites of 3 elementary constituents: left eye, right eye, and mouth.

Extra statistics determine the variable constituent-composition of each face. Distinct expressions are constructed as distinct combinations of distinct eye- and mouth-shapes: left eyes, mouths and right eyes are each selected randomly from 4 alternative shape-variants of smile or 4 alternative shape-variants of frown. The complete alphabet of elementary constituents consists of 4 variants of each smiling or frowning left eye, right eye, or mouth, at each image location ($2\times3\times4\times256\times256$ elementary constituents in total). Faces of various shapes are constructed by varying the relative separations of the eyes and mouth: the displacements (measured from the nominal face-centre in pixel-coordinates $[h, v]$) of the right eye, mouth and left eye each follow Gaussian distributions, about respective means of $[-10,12]$, $[0,-12]$ and $[10,12]$, all with standard deviations of $[\pm3,\pm2]$.

Each S-face is constructed from eyes and mouths each selected independently from among the (4 equally likely) smiling variants with probability 0.8, or from the (4 equally likely) frowning variants with probability 0.2: F-faces contain predominantly frowning features, with the converse probabilities. S and F are statistically distinct classes of face[9]: on average, the members of the same class share more constituents in common than do members of different classes.

[9]The cross-probability of 0.2 represents a fair degree of mixture between the two nominal classes S and F: if this cross-probability were as high as 0.5, no statistical distinction could be drawn between the two classes.

4.5. Using ϑ to select between elementary and composite constituents

The cartoon-faces pattern-ensemble has two kinds of constituent, corresponding to two different defining projections a_i^{o}: elementary facial features and composite faces. (Each composite constituent is a particular combination of 3 specific facial features.) A composite constituent always has a defining projection larger than those of the elementary constituents comprising it [Appendix A.]: measured in units in which whole faces have defining projection a^{o}, the defining projection for the individual facial features is $a^{\mathrm{o}}/\sqrt{3}$, because in the present case 3 orthogonal elementary constituents comprise each composite constituent. We expect [Section 3.6.] the elementary facial features to be the ISO stable points when ϑ is just less than $a^{\mathrm{o}}/\sqrt{3}$, whereas whole faces will be the stable points when ϑ is just less than a^{o}.

Figure 5: The two kinds of ISO convergent synaptic vectors, for the ensemble of patterns made of combinations of *composite* constituents (the variable faces of figure 2). The ISO parameter ϑ is used to select between elementary and composite constituents (on the basis of their different defining projections, respectively $a^{\mathrm{o}}/\sqrt{3}$ and a^{o}). (*a*) When ϑ is set to $0.95\,a^{\mathrm{o}}$, the organised synaptic vectors are detectors for individual faces with specific expressions: *particular* combinations of 3 *specific* facial features. (*b*) When ϑ is $0.95\,a^{\mathrm{o}}/\sqrt{3}$, the convergent states are detectors for (specific) individual facial features. These eye- and mouth-detectors are activated much more frequently than the higher-threshold face-detectors, which are triggered only by particular combinations of two eyes and mouth. (Numerical simulation with path-following from low ϑ.)

Figures 5(*a*) and 5(*b*) respectively show typical ISO convergent vectors at $\vartheta = 0.95\,a^{\mathrm{o}}$ and $\vartheta = 0.95\,a^{\mathrm{o}}/\sqrt{3}$. In both cases ISO finds the constituents that are stable at the corresponding ϑ-value: the elementary constituents at the lower value and the composite constituents at the higher value. At the lower ϑ-value, the adapted neuron responds whenever the pattern contains the specific eye or mouth that matches its receptive field exactly: at the higher ϑ-value, the neuron responds far less frequently, whenever the pattern contains the specific matching conjunction of eyes and mouths.

4.6. Discrimination and generalisation and are antagonistic within the same level

A single neuron can *either* be a discriminating detector for a specific composite constituent (at higher threshold ϑ), *or* a discriminating detector for a specific elementary constituent (at lower threshold). With only one level of processing, the neuron cannot generalise over for a broad class of composite constituents (all the S-faces for example) whilst discriminating between distinct classes (between S and F): within the same processing level, discrimination and generalisation and are antagonistic.

This dilemma is resolved by introducing multiple levels of processing: the separate tasks of discrimination and generalisation can be performed by neurons in separate levels, and are no longer antagonistic.

4.7. Discriminating generalisation in a multiple-level system

A two-level hierarchy of ISO neurons can discriminate between the distinct face-classes S and F whilst generalising over the members within each class.

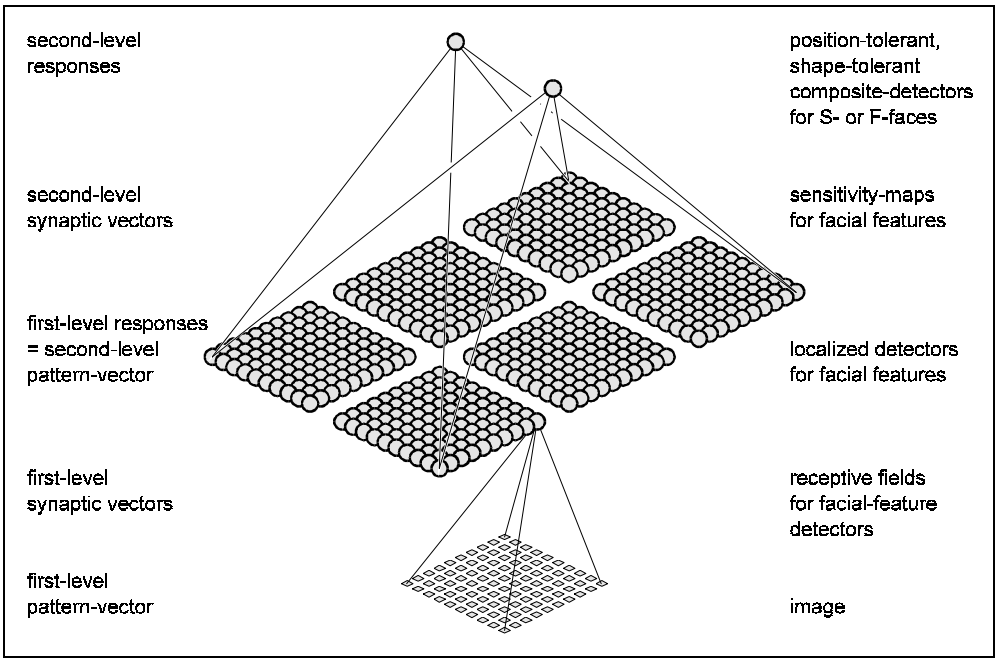

Figure 6: Higher-level neurons can be *generalising* (*i.e.* shape-tolerant, locally position-tolerant) detectors for *similar* composite constituents (similar combinations of elementary constituents). Here, each first-level neuron detects a differently-shaped eye or mouth at a different image-location (detectors for 2×3 out of the 2×3×4 shapes, at 10×10 out of the 256×256 locations, are illustrated here): each square of first-level responses provides a 'map' indicating where in the image-plane each particular kind of feature occurs if present. (The depicted ordering of neurons within levels is purely illustrative: an ISO neuron is oblivious to any ordering of its input-synapses, and infers topographic relationships between its inputs *purely* through learning their statistical interdependences.) The second-level neurons respond to individual faces, *i.e.* to simultaneous activity in triple-combinations of first-level detectors: they generalise over a *variety* of such triple-combinations. Localised receptive fields and tolerant position-localisation emerge through ISO despite complete interconnectivity between levels. In this composite-face example, the first-level detectors have no intrinsic shape- or position-tolerance: in general however, every level of an ISO hierarchy can have a degree of intrinsic transformation-tolerance.

Each of many neurons in the first level adapts directly to the image-space, with ϑ set to $0.95a^{\circ}/\sqrt{3}$, as described in section 4.5. Between them, the adapted first-level neurons form a complete representation of the alphabet of elementary constituents: the $2\times3\times4\times256\times256$ variants of eyes and mouth. Each neuron in the second level receives one

input synapse from every neuron in the first level: each receives the same *second-level* pattern-vector $(r_1,\ldots,r_i,\ldots,r_N)$ made up of the $N = 2\times3\times4\times256\times256$ first-level responses. The resulting network architecture is very simple: complete interconnectivity between layers. How the coding-functions are distributed within this architecture is illustrated in figure 6.

Because 3 of these first-level neurons are activated for each of the n faces appearing in the image (a left-eye detector, a right-eye detector and a mouth detector), any second-level pattern-vector will have only $3n$ out of its $2\times3\times4\times256\times256$ coordinates non-zero. The individual faces are constituents associated with planes of *second-level* pattern-vectors, which cast defining projections $a^{\mathrm{o}}_{(2)}$ (in these units the $3n$ non-zero coordinates each have the value $a^{\mathrm{o}}_{(2)}/\sqrt{3}$, so any second-level pattern-vector encoding an n-face image would have length $\sqrt{n}\,a^{\mathrm{o}}_{(2)}$.

With the first-level neurons encoding the pattern-ensemble in terms of its elementary constituents, the task of expression-invariant discrimination between the two people's faces is made easy for the second-level neurons. They can concentrate on generalisation over expression and face-shape within class S or within class F, because the first level has dealt with the discrimination between smiling and frowning features: two functions that would have been antagonistic within a single level.

Figures 7(*a*) and 7(*b*) are maps depicting two *second-level* synaptic vectors: convergent states of ISO adaptation to the ensemble of second-level pattern-vectors, obtained from two different random initial states at $\vartheta = 0.01\,a^{\mathrm{o}}_{(2)}$. Both these synaptic-weight maps are profiles of a second-level neuron's sensitivity to the various kinds of elementary constituent detected by the first-level neurons: the weights illustrated in figure 7 are ordered as in figure 6. Figures 7(*a*) and 7(*b*) are typical of all the second-level convergent vectors at this ϑ: all the others[10] are just translations of one of these two, to other image locations. These level-two neurons are detectors for individual, whole faces: their thresholds are such that 3 elementary constituents (a left eye, a mouth, and a right eye, at appropriate relative locations) must all be present simultaneously in order to induce a response. Figure 7(*a*) and figure 7(*b*) are discriminating detectors for, respectively, faces of class S or faces of class F: the maps show large positive weights only for smiling features in vector (*a*) and for frowning features in vector (*b*). (Neurons adapting from different random initial states develop into detectors for faces in different image-regions and different classes.)

[10]All $2\times256\times256$ convergent states are equally likely outcomes, from uniformly random initial states.

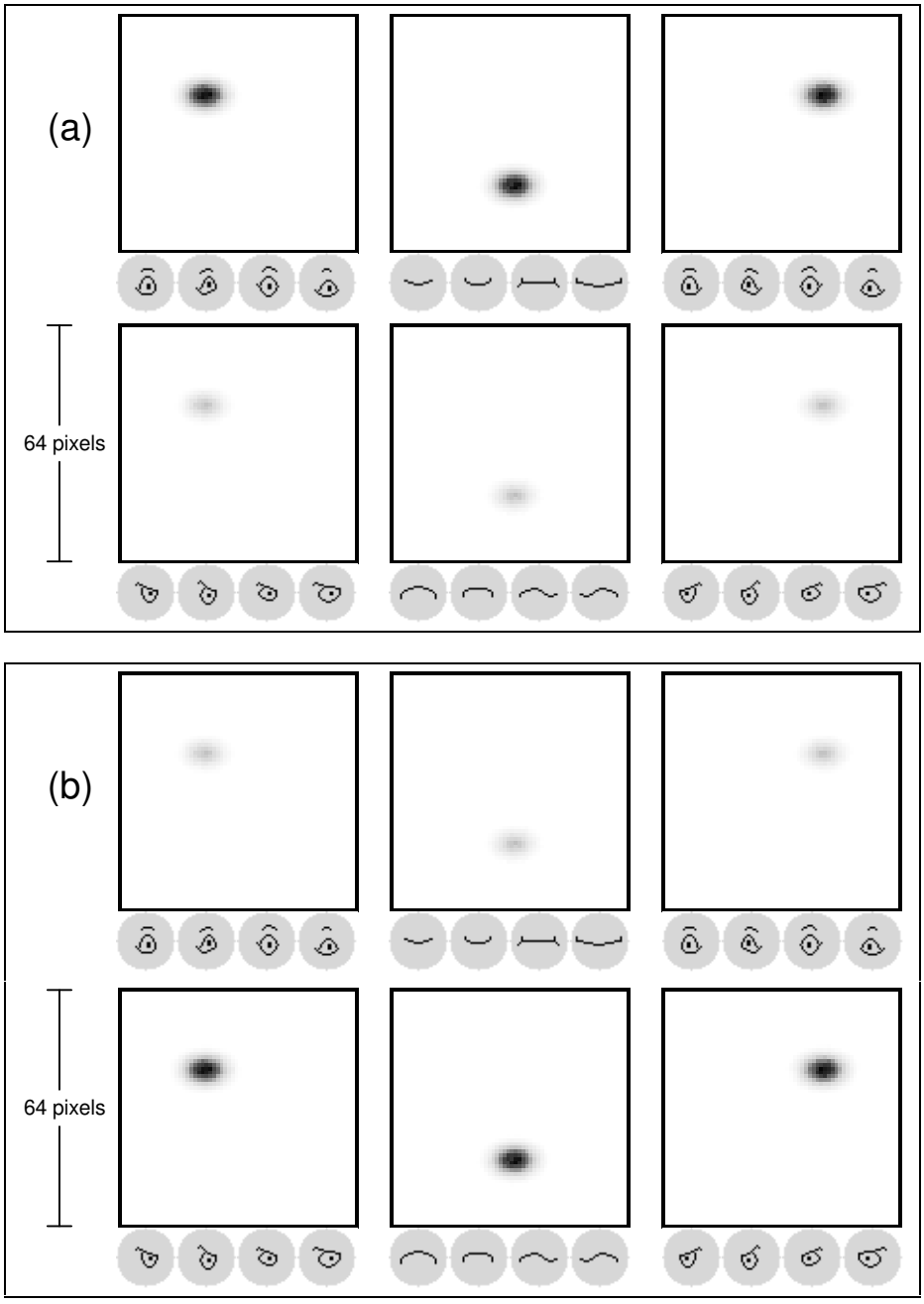

Figure 7: (*a*) and (*b*) are maps of synaptic weights (ordered according to image-location as illustrated in figure 6), for two organised *second-level* synaptic vectors adapted to the variable composite-faces ensemble $(\vartheta = 0.01 a_{(2)}^{o})$. Each square is a profile of the second-level neuron's sensitivity to the location of a specific variant of facial feature (labelled by the attached discs). The converged weight-profiles are identical for the 4 shape-variants of each smiling or frowning eye or mouth, so only one of the 4 profiles for each facial feature is shown. Dark points represent large positive weights, light points less positive weights. One of these two synaptic vectors (*a*) is sensitive to only smiling S faces, the other (*b*) only to frowning F faces. Each neuron is a localised face-detector with shape-tolerance and local position-tolerance. All other convergent states at this ϑ are translated equivalents of one of these two, and are detectors for similar combinations of features localised elsewhere in the image-plane. (Numerical simulations, with path-following from two random lower-ϑ initial states.)

After self-organisation, the sensitivity-profile to each elementary constituent follows a smooth, Gaussian-like dependence on image-location: the second-level neurons are sensitive to each elementary constituent within a region of the image which is localised in position, but *only within a limited degree of precision*. Within these sensitive regions, each sensitivity-profile varies smoothly across neighbouring locations: synapses from first-level detectors for identical facial features at neighbouring locations have approximately equal weights. In other words, second-level pattern-vectors encoding faces that differ only in having their features slightly displaced will cast *equal projections* onto a converged second-level synaptic vector. Consequently, a second-level neuron's response will remain (locally) invariant as any facial feature is moved continuously across its smooth envelope of sensitivity. In addition, equality of weights between the 4 expression-variants of each feature (at each location) means that the response is invariant with respect to the degree of smile or frown. The Gaussian-like sensitivity-profiles have developed to reflect the Gaussian scatter experienced in the facial features' relative separations: the aspect-ratio of each sensitivity-region (3/2), which gives level-two neurons greater tolerance to horizontal feature displacement than to vertical displacement, has adapted to match the aspect-ratio of the distribution of displacements experienced. The equal weights for expression-variants have developed to reflect their equal frequencies of occurrence. Because ϑ is so far below $a^{o}_{(2)}$, the response of each second-level class-detector generalises over a large number of similar faces within its class (if ϑ were only just below $a^{o}_{(2)}$, each neuron would respond to only one specific combination of 3 specific facial features): each class-detector generalises over many distinct faces because its receptive field is *equally* sensitive to facial features of many alternative variants and locations.

Despite their capacity to discriminate between the two face-classes, organised second-level neurons respond invariantly under local face-shifts and transformations of face-shape and expression within each class. These transformation-invariances develop to correspond to geometrical invariances of the pattern-vector distribution: synaptic vectors become oriented normal to planes of pattern-vectors in their response-sectors.

4.8. Local position-tolerance of generalising composite-detectors

The response of a shape-tolerant composite-face detector is locally invariant to deformations in the relative positions of the three elementary features, because the smooth sensitivity-profiles cause near-optimal responses to be achieved when those features are only imprecisely localised. This lack of face-shape precision also gives the composite-detector tolerance to local shifts of the face's position: the response remains locally invariant as the face's nominal location is moved continuously across the detector's sensitive region.

This locally shift-invariant composite-detector defines a face's position with only a limited degree of precision, which matches the extent of its shape-tolerance. This is desirable, because it would be pointless to perceive a thing's location with a precision much finer than the precision with which its shape is perceived [Webber 1991].

In multiple-level vision systems, ISO can generate shape-tolerant and locally position-tolerant composite-detectors at each level, which have tolerances matched to the shape-variation experienced within the composites detected by that level. Through self-organisation of invariances that are data-driven rather than enforced *a priori*, a nested hierarchy of localised receptive fields emerges, whose variation-tolerances are similar to those designed into the Neocognitron model of tolerant recognition [Fukushima 1982].

4.9. Representation of constituent-continua: generalisation over continuous transformations

In the cartoon-faces ensemble, the elementary constituents were discrete: the eyes and mouths came in just 2×3×4 distinct shape-variants, and a well-defined number of elementary-constituent detectors could encode this discrete set. The transformations that generated the variation in these constituents were discrete transformations. ISO constituent-detectors are equally capable of representing a continuum of constituent-variants, and of generalising over continuous transformations of constituents: this is illustrated both by section 4.3.'s example of (single-level) generalisation over local variation in face-position and by section 4.7.'s example of (second-level) generalisation over variation in the absolute and relative positions of facial features.

A continuum of constituents is encoded in just the same way as a discrete set, though each distinct constituent-detector now generalises over a distinct sub-range of closely similar variants within the continuum, rather than detecting just one discrete variant. The detector's ϑ-parameter controls how broad or narrow this sub-range is, as in sections 4.3. and 4.7.

Continuous transformations are transformations that can be generated by repeated application of infinitesimal transformations. Visual examples of continuous constituent-transformations include position-shift, rotation, scaling, stereo-disparity change, and arbitrary shape-deformation: the infinitesimal generators of these transformations involve displacements within the image that are small compared to the constituent's size.

ISO can represent constituent-variants under *any* kind of continuous transformation in this way. This follows because any infinitesimal transformation on a pattern-constituent is equivalent to an infinitesimal displacement of the pattern-vector within its pattern-space: consequently, pattern-vectors generated from one another by small but finite continuous constituent-transformations must lie within a *localised volume* of the pattern-space. Therefore, a plane of pattern-vectors will retain its coplanarity, to some degree of approximation, under these small constituent-transformations: the normal-vectors of these approximate planes will be convergent states of ISO synaptic vectors, which will each detect small sub-ranges of the constituent-continuum and generalise within those ranges. ISO constituent-detectors also generalise over additive noise for the same reason: additive noise transforms a pattern-vector within a localised volume of the pattern-space.

5. Conclusion

Wherever pattern-vectors are concentrated on planes, invariance of pattern-response can self-organise. In this way, individual neurons can develop into detectors for the individual constituent-structures which make up the various patterns: their responses become invariant with respect to other pattern-attributes. Such response-invariances can take the form of generalisation over alternative constituents: the same response whichever of the alternatives happens to be present. Patterns' constituents can themselves be composites of yet more elementary constituents: in multiple-level systems, composite-detectors can generalise over the variation in that composition. Generalisation over constituent-composition is a general class of transformation-invariance, which includes the visual transformation-invariances of shape-tolerance and local position-tolerance.

Appendices

A. Composite constituents

Consider m coplanar pattern-subsets $1,2,\ldots,m$ of an ensemble, casting projections $a_1^{\mathrm{o}},a_2^{\mathrm{o}},\ldots,a_m^{\mathrm{o}}$ along (linearly independent) elementary constituents $\hat{\boldsymbol{c}}_1,\hat{\boldsymbol{c}}_2,\ldots,\hat{\boldsymbol{c}}_m$. Patterns of the m subsets lie on the m planes

$$\boldsymbol{x}\cdot\hat{\boldsymbol{c}}_i \;=\; a_i^{\mathrm{o}} \qquad\qquad (i=1,2,\ldots,m)\ . \tag{28}$$

Patterns containing the conjunction of all m constituents lie in the intersection of the m planes. This intersection contains, and is normal to, the vector

$$\boldsymbol{v} \;=\; v_1\hat{\boldsymbol{c}}_1+v_2\hat{\boldsymbol{c}}_2+\cdots+v_m\hat{\boldsymbol{c}}_m\ , \tag{29}$$

where

$$v_j \;=\; \sum_{i=1}^{m}\left(\boldsymbol{M}^{-1}\right)_{ji}a_i^{\mathrm{o}} \qquad\qquad \left(\boldsymbol{M}_{ji}\;\equiv\;\hat{\boldsymbol{c}}_j\cdot\hat{\boldsymbol{c}}_i\right)\ . \tag{30}$$

The vector $\boldsymbol{v}$ must lie in the intersection and be normal to it: inversion of (30) shows that $\boldsymbol{v}$ lies in all m planes (28). The unit vector

$$\hat{\boldsymbol{v}} \;\equiv\; |\boldsymbol{v}|^{-1}\boldsymbol{v} \tag{31}$$

is a constituent normal to the coplanar intersection of subsets $1,2,\ldots,m$: it is a composite of the more elementary constituents $\hat{\boldsymbol{c}}_1,\hat{\boldsymbol{c}}_2,\ldots,\hat{\boldsymbol{c}}_m$. All patterns containing the conjunction of $\hat{\boldsymbol{c}}_1,\hat{\boldsymbol{c}}_2,\ldots,\hat{\boldsymbol{c}}_m$ cast a projection $|\boldsymbol{v}|$ along $\hat{\boldsymbol{v}}$. It is easy to see (by construction) that the projection associated with the composite is larger than any of the projections associated with the elementary constituents:

$$|\boldsymbol{v}| \;>\; a_i^{\mathrm{o}} \qquad\qquad \text{for all } i=1,2,\ldots,m\ . \tag{32}$$

All the patterns containing the conjunction of $\hat{\boldsymbol{c}}_1,\hat{\boldsymbol{c}}_2,\ldots,\hat{\boldsymbol{c}}_m$ can be segregated from other patterns by a threshold-cut ϑ (in their projection along $\hat{\boldsymbol{v}}$) which is just less than $|\boldsymbol{v}|$.

B. Constraining normalisation in the ISO dynamics

An independent normalisation process can be blended into the ISO dynamics (21) explicitly, by adding a constraint term to the dynamical equation:

$$\frac{d\boldsymbol{w}}{dt} \;=\; \lambda\left\langle \boldsymbol{x}\,\theta(\boldsymbol{w}\cdot\boldsymbol{x}-\vartheta|\boldsymbol{w}|)-\phi\boldsymbol{w}\right\rangle_{\{\boldsymbol{x}\}} \qquad\qquad (\text{scalar } \phi(\boldsymbol{x},\boldsymbol{w}))\ . \tag{33}$$

Because the constraint term $\phi\boldsymbol{w}$ is always parallel (or antiparallel) to the direction of $\boldsymbol{w}$, its only effect is continuous rescaling of $\boldsymbol{w}$, and it cannot interfere with the prescription (21) for the adaptation of $\boldsymbol{w}$'s direction. If one sets

$$\phi \;=\; |\boldsymbol{w}|^{-2}(\boldsymbol{w}\cdot\boldsymbol{x})\,\theta(\boldsymbol{w}\cdot\boldsymbol{x}-\vartheta|\boldsymbol{w}|)\ , \tag{34}$$

the normalisation of $\boldsymbol{w}$ will be constrained constant, since then

$$\frac{d}{dt}|\boldsymbol{w}|^2 \;=\; 2\boldsymbol{w}\cdot\frac{d\boldsymbol{w}}{dt} \;=\; 0\ . \tag{35}$$

C. ISO as gradient optimisation

The ISO dynamics can be derived as a gradient ascent

$$\frac{d\boldsymbol{w}}{dt} \;=\; \lambda' \frac{\partial V}{\partial \boldsymbol{w}} \tag{36}$$

on a potential

$$V(\boldsymbol{w}) \;\equiv\; \tfrac{1}{2}\Big\langle \theta\big(\hat{\boldsymbol{w}}\cdot\boldsymbol{x} - \vartheta\big)^2 \Big\rangle_{\{\boldsymbol{x}\}} \;. \tag{37}$$

(36) prescribes a dynamical equation

$$\frac{d\boldsymbol{w}}{dt} \;=\; \lambda' \,|\boldsymbol{w}|^{-2} \Big\langle \theta\big(\boldsymbol{w}\cdot\boldsymbol{x} - \vartheta\,|\boldsymbol{w}|\big)\,\big[\boldsymbol{x} - |\boldsymbol{w}|^{-2}(\boldsymbol{w}\cdot\boldsymbol{x})\,\boldsymbol{w}\big] \Big\rangle_{\{\boldsymbol{x}\}} \;. \tag{38}$$

This always leaves $|\boldsymbol{w}|$ constant in time, because it guarantees (35), and so one can define another constant

$$\lambda \;\equiv\; \lambda'\,|\boldsymbol{w}|^{-2} \;. \tag{39}$$

Thus, the dynamical equation (33),(34) of self-normalising ISO is recovered from gradient ascent with respect to $\boldsymbol{w}$.

ISO also performs gradient ascent in the *normalised* synaptic vector $\hat{\boldsymbol{w}}$, as well as in $\boldsymbol{w}$ itself, as can be seen by deriving the gradient of V with respect to $\hat{\boldsymbol{w}}$, subject to the constraint

$$|\hat{\boldsymbol{w}}|^2 \;=\; 1 \;. \tag{40}$$

The projection of this gradient that lies on the surface (40) is

$$\left.\frac{\partial V}{\partial \hat{\boldsymbol{w}}}\right|_{|\hat{\boldsymbol{w}}|} \;=\; \Big\langle \boldsymbol{x}\,\theta\big(\hat{\boldsymbol{w}}\cdot\boldsymbol{x} - \vartheta\big) - \varphi\hat{\boldsymbol{w}} \Big\rangle_{\{\boldsymbol{x}\}} \qquad\qquad \big(\text{scalar } \varphi(\boldsymbol{x},\hat{\boldsymbol{w}})\big) \;, \tag{41}$$

where the Lagrange multiplier φ is set to

$$\varphi \;=\; (\hat{\boldsymbol{w}}\cdot\boldsymbol{x})\,\theta\big(\hat{\boldsymbol{w}}\cdot\boldsymbol{x} - \vartheta\big) \tag{42}$$

in order to satisfy (40) and

$$2\hat{\boldsymbol{w}}\cdot\frac{d\hat{\boldsymbol{w}}}{dt} \;=\; \frac{d}{dt}|\hat{\boldsymbol{w}}|^2 \;=\; 0 \;. \tag{43}$$

Gradient ascent in $\hat{\boldsymbol{w}}$ would therefore satisfy

$$\frac{d\hat{\boldsymbol{w}}}{dt} \;=\; \lambda \Big\langle \big[\boldsymbol{x} - (\hat{\boldsymbol{w}}\cdot\boldsymbol{x})\hat{\boldsymbol{w}}\big]\,\theta\big(\hat{\boldsymbol{w}}\cdot\boldsymbol{x} - \vartheta\big) \Big\rangle_{\{\boldsymbol{x}\}} \;. \tag{44}$$

This prescribes the *same* trajectory as the self-normalising ISO dynamics (33),(34) would prescribe, for any $\boldsymbol{w}$ on the surface

$$|\boldsymbol{w}|^2 \;=\; 1 \;. \tag{45}$$

Homogeneity of the dynamical equation means that all ISO trajectories are mapped *invariantly* onto this surface by scale-transformation (22). Thus, ISO is equivalent to gradient ascent in V with respect to the normalised synaptic vector $\hat{\boldsymbol{w}}$.

D. Inevitability of convergence

The normalised synaptic vector $\hat{\boldsymbol{w}}$ necessarily converges under the ISO dynamics.

From the definition (37) of the potential $V(\boldsymbol{w})$, it is clear that it must always have a finite upper bound, which is exceeded for no value of $\boldsymbol{w}$ (assuming all $\boldsymbol{x}$ in $\{\boldsymbol{x}\}$ have finite

norm). V must therefore have only finite maxima over $\hat{\boldsymbol{w}}$. Since ISO performs gradient ascent in $\hat{\boldsymbol{w}}$ on V [Appendix C.], the ISO dynamics must *always* converge in $\hat{\boldsymbol{w}}$, to a maximum of V. Stability of the convergent state is guaranteed by Lyapunov's first theorem.

The inevitability of ISO convergence has one caveat. ISO can be stationary at $\hat{\boldsymbol{w}}$

$$\frac{d\hat{\boldsymbol{w}}}{dt} = \boldsymbol{0} \tag{46}$$

without convergence (stranded), if no pattern in $\{\boldsymbol{x}\}$ lies within $\boldsymbol{w}$'s response-sector, *i.e.*

$$\theta\big(\hat{\boldsymbol{w}}\cdot\boldsymbol{x} - \vartheta\big) = 0 \qquad \text{for all } \boldsymbol{x} \text{ in } \{\boldsymbol{x}\} \ . \tag{47}$$

This implies

$$V(\boldsymbol{w}) = 0 \ , \tag{48}$$

as is clear from V's definition (37). Also clear is that V can never be less than zero for any value of $\boldsymbol{w}$. Since the gradient ascent can never allow V to decrease, the stranded condition (48) can never be reached from an unstranded state

$$V(\boldsymbol{w}) > 0 \ . \tag{49}$$

Provided the initial state $\boldsymbol{w}$ has at least one pattern-vector above threshold at the outset, the ISO dynamics can never *subsequently* become stranded, and must inevitably converge to a maximum of V.

E. General conditions for ISO convergent states

The general condition for a normalised synaptic vector $\hat{\boldsymbol{w}}^{\infty}$ to be a stationary point (25) of the dynamics (23) is

$$\Big\langle \boldsymbol{x}\,\theta\big(\hat{\boldsymbol{w}}^{\infty}\cdot\boldsymbol{x} - \vartheta\big)\Big\rangle_{\{\boldsymbol{x}\}} = b\hat{\boldsymbol{w}}^{\infty}, \qquad \text{scalar } b = \Big\langle (\hat{\boldsymbol{w}}^{\infty}\cdot\boldsymbol{x})\,\theta\big(\hat{\boldsymbol{w}}^{\infty}\cdot\boldsymbol{x} - \vartheta\big)\Big\rangle_{\{\boldsymbol{x}\}} \tag{50}$$

or

$$\big\langle (\hat{\boldsymbol{w}}^{\infty}\cdot\boldsymbol{x})\boldsymbol{x}\big\rangle_{\Theta} - \vartheta\langle\boldsymbol{x}\rangle_{\Theta} = b\hat{\boldsymbol{w}}^{\infty}, \qquad b = \big\langle (\hat{\boldsymbol{w}}^{\infty}\cdot\boldsymbol{x})^2\big\rangle_{\Theta} - \vartheta\,\hat{\boldsymbol{w}}^{\infty}\cdot\langle\boldsymbol{x}\rangle_{\Theta} \ , \tag{51}$$

where taking the ensemble-average just over the subset $\Theta(\hat{\boldsymbol{w}},\vartheta)$ of patterns in the response-sector (18) allows the non-linear function $\theta(\zeta)$ to be dropped.

Stability is demonstrated by considering small perturbations

$$\hat{\boldsymbol{w}} = \hat{\boldsymbol{w}}^{\infty} + \boldsymbol{\varepsilon} \tag{52}$$

away from the stationary point, for which

$$\hat{\boldsymbol{w}}^{\infty}\cdot\boldsymbol{\varepsilon} = 0 \tag{53}$$

to first order in $\boldsymbol{\varepsilon}$, because $|\hat{\boldsymbol{w}}| = |\hat{\boldsymbol{w}}^{\infty}|$. For any finite ensemble of patterns, $\hat{\boldsymbol{w}}$'s response-sector will contain the same patterns Θ as $\hat{\boldsymbol{w}}^{\infty}$'s response-sector, if $\boldsymbol{\varepsilon}$ is small enough:

$$\Theta(\hat{\boldsymbol{w}},\vartheta) = \Theta(\hat{\boldsymbol{w}}^{\infty},\vartheta) \equiv \Theta \ . \tag{54}$$

(The vicinity $\{\hat{\boldsymbol{w}}\}$, for which $\hat{\boldsymbol{w}}$'s response-sector contains the same patterns Θ, can often be large, depending on the distribution of patterns.) Within this vicinity, the ensemble-average in the dynamical equation (23) can be taken just over Θ:

$$\frac{d\hat{\boldsymbol{w}}}{\lambda dt} = \left\langle (\hat{\boldsymbol{w}}\cdot\boldsymbol{x}-\vartheta)[\boldsymbol{x}-(\hat{\boldsymbol{w}}\cdot\boldsymbol{x})\hat{\boldsymbol{w}}]\right\rangle_{\Theta} . \tag{55}$$

It follows from (51) and (53) that

$$\left\langle (\boldsymbol{\varepsilon}\cdot\boldsymbol{x})(\hat{\boldsymbol{w}}^{\infty}\cdot\boldsymbol{x})\right\rangle_{\Theta} = \vartheta\langle\boldsymbol{x}\rangle_{\Theta}\cdot\boldsymbol{\varepsilon} . \tag{56}$$

Substituting (52) for the perturbed synaptic vector, and using (51) and (56), the velocity-field near the stationary point can be expanded to first order in $\boldsymbol{\varepsilon}$:

$$\frac{d\hat{\boldsymbol{w}}}{\lambda dt} = \left\langle (\boldsymbol{\varepsilon}\cdot\boldsymbol{x})\boldsymbol{x}\right\rangle_{\Theta} + \vartheta(\langle\boldsymbol{x}\rangle_{\Theta}\cdot\hat{\boldsymbol{w}}^{\infty})\boldsymbol{\varepsilon} - \left\langle (\hat{\boldsymbol{w}}^{\infty}\cdot\boldsymbol{x})^2\right\rangle_{\Theta}\boldsymbol{\varepsilon} - \vartheta\left(\langle\boldsymbol{x}\rangle_{\Theta}\cdot\boldsymbol{\varepsilon}\right)\hat{\boldsymbol{w}}^{\infty} + O(\boldsymbol{\varepsilon}^2) . \tag{57}$$

The condition for the stationary point to be stable is

$$\frac{d}{dt}|\boldsymbol{\varepsilon}|^2 = 2\boldsymbol{\varepsilon}\cdot\frac{d\hat{\boldsymbol{w}}}{dt} < 0 \tag{58}$$

or

$$\left\langle (\hat{\boldsymbol{\varepsilon}}\cdot\boldsymbol{x})^2\right\rangle_{\Theta} < \left\langle (\hat{\boldsymbol{w}}^{\infty}\cdot\boldsymbol{x})^2\right\rangle_{\Theta} - \vartheta\langle\boldsymbol{x}\rangle_{\Theta}\cdot\hat{\boldsymbol{w}}^{\infty} \qquad (\hat{\boldsymbol{\varepsilon}}\equiv|\boldsymbol{\varepsilon}|^{-1}\boldsymbol{\varepsilon}) \tag{59}$$

for all small perturbations $\boldsymbol{\varepsilon}$ satisfying (53).

Within the vicinity $\{\hat{\boldsymbol{w}}\}$ for which $\hat{\boldsymbol{w}}$'s response-sector contains the same patterns Θ, the potential $V(\hat{\boldsymbol{w}})$ is quadratic, so $\{\hat{\boldsymbol{w}}\}$ can contain one or two stable convergent states, or none (or a continuum of metastable convergent states).

F. Conditions for a constituent to be an ISO convergent state

Suppose the subset Θ of patterns that lie on the plane $\boldsymbol{x}\cdot\hat{\boldsymbol{c}}_i = a_i^{\mathrm{o}}$ has its mean aligned with the constituent $\hat{\boldsymbol{c}}_i$:

$$\langle\boldsymbol{x}\rangle_{\Theta} = a_i^{\mathrm{o}}\hat{\boldsymbol{c}}_i . \tag{60}$$

Consider an ISO stationary point $\hat{\boldsymbol{w}}^{\infty}$ aligned with $\hat{\boldsymbol{c}}_i$. Suppose the ISO parameter ϑ is chosen less than a_i^{o}, so that the threshold-plane $\boldsymbol{x}\cdot\hat{\boldsymbol{c}}_i = \vartheta$ separates that subset of patterns from the rest of the ensemble: $\theta(\hat{\boldsymbol{w}}^{\infty}\cdot\boldsymbol{x}-\vartheta)$ is non-zero only for the patterns Θ satisfying $\boldsymbol{x}\cdot\hat{\boldsymbol{c}}_i = a_i^{\mathrm{o}}$. $\hat{\boldsymbol{w}}^{\infty}$ will be a stationary point (51) because

$$\hat{\boldsymbol{w}}^{\infty}\cdot\boldsymbol{x} = a_i^{\mathrm{o}} \qquad (\boldsymbol{x}\text{ in }\Theta) \tag{61}$$

and

$$\langle\boldsymbol{x}\rangle_{\Theta} = a_i^{\mathrm{o}}\hat{\boldsymbol{w}}^{\infty} . \tag{62}$$

The condition (59) that this stationary point $\hat{\boldsymbol{w}}^{\infty} = \hat{\boldsymbol{c}}_i$ is stable takes a simple form. The largest value its left-hand side

$$\left\langle (\hat{\boldsymbol{\varepsilon}}\cdot\boldsymbol{x})^2\right\rangle_{\Theta} = \left\langle (\hat{\boldsymbol{\varepsilon}}\cdot\boldsymbol{x})^2\right\rangle_{\Theta} - \left[\langle\boldsymbol{x}\rangle_{\Theta}\cdot\hat{\boldsymbol{\varepsilon}}\right]^2 = \hat{\boldsymbol{\varepsilon}}^{\mathrm{T}}\left[\langle\boldsymbol{x}\boldsymbol{x}^{\mathrm{T}}\rangle_{\Theta} - \langle\boldsymbol{x}\rangle_{\Theta}\langle\boldsymbol{x}\rangle_{\Theta}^{\mathrm{T}}\right]\hat{\boldsymbol{\varepsilon}} = \hat{\boldsymbol{\varepsilon}}^{\mathrm{T}}\boldsymbol{C}_{\Theta}\hat{\boldsymbol{\varepsilon}} \tag{63}$$

can take (for any $\boldsymbol{\varepsilon}$) is the principal eigenvalue v_{Θ}^{1} of the covariance matrix $\boldsymbol{C}_{\Theta}$ of the coplanar subset Θ. The constituent $\hat{\boldsymbol{c}}_i$ will therefore be a stable convergent state if Θ's principal variance satisfies

$$v_{\Theta}^{1} < a_i^{\mathrm{o}}(a_i^{\mathrm{o}}-\vartheta), \qquad i.e. \qquad \vartheta < a_i^{\mathrm{o}} - v_{\Theta}^{1}/a_i^{\mathrm{o}} . \tag{64}$$

As ϑ is increased towards a_i^{o}, a phase-transition is passed, beyond which the stationary state is no longer stable. The convergent state will bifurcate, along the principal eigenvector of $\boldsymbol{C}_\Theta$.

A constituent $\hat{\boldsymbol{c}}_i$ will be an ISO convergent state if the coplanar patterns which define it have their mean lying along $\hat{\boldsymbol{c}}_i$ and if their principal variance satisfies (64).

G. Approximately coplanar patterns give rise to similar convergent states

An approximately coplanar pattern-subset will give rise to an ISO convergent state in the same way: if exactly coplanar patterns Θ are perturbed away from their plane, the corresponding ISO stable state $\hat{\boldsymbol{w}}^\infty$ will be perturbed away from the location (62). The stable state will move continuously as the patterns become gradually less coplanar, until the stability-condition (59) ceases to be valid.

This is demonstrated by deriving a differential equation that describes the variation of $\hat{\boldsymbol{w}}^\infty$ as the distribution of patterns changes. Within the vicinity $\{\hat{\boldsymbol{w}}\}$ for which $\hat{\boldsymbol{w}}$'s response-sector contains the same patterns Θ, the dynamics (55) depend on two sufficient statistics of those patterns: the matrices $\langle \boldsymbol{x} \rangle_\Theta$ and $\langle \boldsymbol{x}\boldsymbol{x}^{\mathrm{T}} \rangle_\Theta$. The general condition (51) for $\hat{\boldsymbol{w}}^\infty$ to be a stationary point can be written in terms of these matrices:

$$\langle \boldsymbol{x}\boldsymbol{x}^{\mathrm{T}} \rangle_\Theta \hat{\boldsymbol{w}}^\infty - \vartheta \langle \boldsymbol{x} \rangle_\Theta = \beta \hat{\boldsymbol{w}}^\infty \qquad \left(\text{scalar } \beta = \hat{\boldsymbol{w}}^{\infty\mathrm{T}} \langle \boldsymbol{x}\boldsymbol{x}^{\mathrm{T}} \rangle_\Theta \hat{\boldsymbol{w}}^\infty - \vartheta \hat{\boldsymbol{w}}^{\infty\mathrm{T}} \langle \boldsymbol{x} \rangle_\Theta \right). \quad (65)$$

Consider how $\hat{\boldsymbol{w}}^\infty$ moves as the distribution of patterns in Θ is perturbed so that

$$\langle \boldsymbol{x} \rangle_\Theta \to \langle \boldsymbol{x} \rangle_\Theta + \delta \langle \boldsymbol{x} \rangle_\Theta \qquad \text{and} \qquad \langle \boldsymbol{x}\boldsymbol{x}^{\mathrm{T}} \rangle_\Theta \to \langle \boldsymbol{x}\boldsymbol{x}^{\mathrm{T}} \rangle_\Theta + \delta \langle \boldsymbol{x}\boldsymbol{x}^{\mathrm{T}} \rangle_\Theta . \quad (66)$$

The matrices $\langle \boldsymbol{x} \rangle_\Theta$ and $\langle \boldsymbol{x}\boldsymbol{x}^{\mathrm{T}} \rangle_\Theta$ change continuously as Θ's distribution is perturbed continuously from coplanarity: they change by $P_{\boldsymbol{x}}(\boldsymbol{x}_k)\,\delta\boldsymbol{x}_k$ and $P_{\boldsymbol{x}}(\boldsymbol{x}_k)\left[\boldsymbol{x}_k \delta\boldsymbol{x}_k^{\mathrm{T}} + \delta\boldsymbol{x}_k \boldsymbol{x}_k^{\mathrm{T}}\right]$ respectively as each individual pattern-vector $\boldsymbol{x}_k$ is moved by $\delta\boldsymbol{x}_k$. The stationary point will be perturbed to a new position $\hat{\boldsymbol{w}}^\infty + \delta\hat{\boldsymbol{w}}^\infty$ which satisfies

$$\left[\langle \boldsymbol{x}\boldsymbol{x}^{\mathrm{T}} \rangle_\Theta + \delta \langle \boldsymbol{x}\boldsymbol{x}^{\mathrm{T}} \rangle_\Theta\right]\left[\hat{\boldsymbol{w}}^\infty + \delta\hat{\boldsymbol{w}}^\infty\right] - \vartheta\left[\langle \boldsymbol{x} \rangle_\Theta + \delta \langle \boldsymbol{x} \rangle_\Theta\right] = (\beta + \delta\beta)\left[\hat{\boldsymbol{w}}^\infty + \delta\hat{\boldsymbol{w}}^\infty\right], \quad (67)$$

i.e.

$$\langle \boldsymbol{x}\boldsymbol{x}^{\mathrm{T}} \rangle_\Theta \delta\hat{\boldsymbol{w}}^\infty + \delta \langle \boldsymbol{x}\boldsymbol{x}^{\mathrm{T}} \rangle_\Theta \hat{\boldsymbol{w}}^\infty - \vartheta\, \delta \langle \boldsymbol{x} \rangle_\Theta = \beta\, \delta\hat{\boldsymbol{w}}^\infty + \delta\beta\, \hat{\boldsymbol{w}}^\infty . \quad (68)$$

Using (40) and (43), one can infer

$$\delta\beta = \hat{\boldsymbol{w}}^{\infty\mathrm{T}} \langle \boldsymbol{x}\boldsymbol{x}^{\mathrm{T}} \rangle_\Theta \delta\hat{\boldsymbol{w}}^\infty + \hat{\boldsymbol{w}}^{\infty\mathrm{T}} \delta \langle \boldsymbol{x}\boldsymbol{x}^{\mathrm{T}} \rangle_\Theta \hat{\boldsymbol{w}}^\infty - \vartheta\, \hat{\boldsymbol{w}}^{\infty\mathrm{T}} \delta \langle \boldsymbol{x} \rangle_\Theta \quad (69)$$

and so eliminate $\delta\beta$, obtaining the differential equation which relates $\delta\hat{\boldsymbol{w}}^\infty$ to $\delta\langle \boldsymbol{x} \rangle_\Theta$ and $\delta\langle \boldsymbol{x}\boldsymbol{x}^{\mathrm{T}} \rangle_\Theta$:

$$\begin{aligned} \left[\langle \boldsymbol{x}\boldsymbol{x}^{\mathrm{T}} \rangle_\Theta - \hat{\boldsymbol{w}}^\infty \hat{\boldsymbol{w}}^{\infty\mathrm{T}} \langle \boldsymbol{x}\boldsymbol{x}^{\mathrm{T}} \rangle_\Theta - \beta\right] \delta\hat{\boldsymbol{w}}^\infty = \\ \vartheta\, \delta \langle \boldsymbol{x} \rangle_\Theta - \delta \langle \boldsymbol{x}\boldsymbol{x}^{\mathrm{T}} \rangle_\Theta \hat{\boldsymbol{w}}^\infty - \hat{\boldsymbol{w}}^{\infty\mathrm{T}} \left[\vartheta\, \delta \langle \boldsymbol{x} \rangle_\Theta - \delta \langle \boldsymbol{x}\boldsymbol{x}^{\mathrm{T}} \rangle_\Theta \hat{\boldsymbol{w}}^\infty\right] \hat{\boldsymbol{w}}^\infty \end{aligned} \quad (70)$$

Acknowledgements: The author is very grateful to Dr. Steve Luttrell for discussions, to Dr. Sam Hill and Dr. Adar Pelah for opinions on the manuscript, and to Dr. Adar Pelah for making available digitised images. This work was initially funded by the UK Science and Engineering Research Council, but supported to completion by the UK Defence Research Agency.

References

[Barlow 1959] *Sensory mechanisms, the reduction of redundancy, and intelligence*, National Physical Laboratory Symposium on Mechanization of Thought Processes **10**, 535-539, HMSO London.

[Barlow 1972] *Single units and sensation: a neuron doctrine for perceptual psychology?*, Perception **1**, 371-394.

[Field 1987] *Relations between the statistics of natural images and the response properties of cortical cells*, Journal of the Optical Society of America **4**, 2379-2394.

[Field 1993] *Scale-invariance and self-similar `wavelet' transforms: An analysis of natural scenes and mammalian visual systems*, in: Wavelets, Fractals and Fourier Transforms: New Developments and New Applications, eds. Farge, Hunt & Vassilicos, Oxford University Press.

[Hebb 1949] *The Organisation of Behaviour*, Wiley.

[Fukushima & Miyake 1982] *Neocognitron: A new algorithm for pattern recognition tolerant of deformations and shifts in position*, Pattern Recognition **15**, 455-469.

[Kohonen 1989] *Self-organisation and Associative Memory* (third edition), Springer.

[Ljung 1977] *Analysis of recursive stochastic algorithms*, IEEE Transactions on Automatic Control **AC-22**, 551-575.

[Minsky & Papert 1969] *Perceptrons*, MIT.

[Oja 1982] *A simplified neuron model as a principal component analyser*, Journal of Mathematical Biology **15**, 267-273.

[Rumelhart & Zipser 1985] *Feature discovery by competitive learning*, Cognitive Science **9**, 75-112.

[Shannon 1949] *The Mathematical Theory of Communication*, Illinois University.

[Stewart 1992] *Bifurcation theory old and new*, in: The Dynamics of Numerics and the Numerics of Dynamics, eds. Broomhead & Iserles, IMA Conference Series **34**, 31-67, Clarendon.

[Webber 1991] *Self-organisation of position- and deformation-tolerant neural representations*, Network: Computation in Neural Systems **1**, 43-61.