The Sparse Manifold Transform

Bruno Olshausen
Redwood Center for Theoretical Neuroscience,
Helen Wills Neuroscience Institute, and School of Optometry
UC Berkeley

with Yubei Chen (EECS) and Dylan Paiton (Vision Science)

NSF 1718991 RI: Extracting and understanding sparse structure in spatiotemporal data, PI: Sommer
Are there principles?

“God is a hacker”
– Francis Crick

“Individual nerve cells were formerly thought to be unreliable… This was quite wrong, and we now realise their apparently erratic behavior was caused by our ignorance, not the neuron’s incompetence.”
– H.B. Barlow (1972)
faces

'Gabor filters' . . . ? . . . objects . . . faces

an absolute depth judgment with respect to fixation, while fine stereopsis requires the judgment of relative depth, i.e., comparing depth across space; (2) the particular coarse stereopsis task used requires the monkey to discriminate a signal in noise, while the fine task does not; (3) the range of disparities is quite different. Chowdhury and DeAngelis (2008) replicate the finding that monkeys initially trained on coarse stereopsis show impaired coarse depth discrimination when muscimol is injected into MT. Remarkably, the same animals, after a second round of training on fine stereopsis, are unimpaired at either fine or coarse depth discrimination by similar injections. Moreover, recordings in MT show that neuronal responses are not altered by learning the fine stereopsis task. Given the differences between the tasks and the large number of visual areas containing disparity-sensitive neurons, one might not be surprised to find different areas involved in the two tasks. But it is quite unexpected that merely learning one task would change the contribution of areas previously involved in the other. Chowdhury and DeAngelis conclude that the change in outcome reflects a change in neural decoding—decision centers that decode signals to render judgments of depth, finding MT signals unreliable for the fine stereopsis task, switch their inputs to select some better source of disparity information. Candidates include ventral stream areas V4 or IT, where relative disparity signals have been reported (Orban, 2008) and which contain far more neurons (Figure 1).

Chowdhury and DeAngelis' monkeys were trained simultaneously or previously to discriminate motion, which engages MT. Faced with a qualitatively similar random dot stimulus, it might make sense for the cortex to try to solve the new problem of stereopsis with existing decoding strategies. But if the animals were initially trained on a different task—say, a texture discrimination—MT might never be engaged at all. It would also be interesting to see the outcome if monkeys were trained on depth tasks that were less different and could be interleaved in the same sessions, for example noise-limited depth judgments using similar absolute or relative disparity
Three principles of unsupervised learning

Sparse coding $\rightarrow$ feature selectivity

Manifold flattening $\rightarrow$ equivariance

Persistence $\rightarrow$ invariance
Sparse coding
The ‘Ratio Club’ (1952)

Harald Shipton, John Bates, W.R. Hick, John Pringle, Donald Shell, John Westcott, Donald Mackay.
Giles Brindley, Tom McLardy, Ross Ashby, Thomas Gold, Albert Utley.
Alan Turing, Surney Sutton, William Rushton, George Dawson, Horace Barlow.
Single units and sensation: A neuron doctrine for perceptual psychology?

H B Barlow
Department of Physiology-Anatomy, University of California, Berkeley, California 94720
Received 6 December 1972

Abstract. The problem discussed is the relationship between the firing of single neurons in sensory pathways and subjectively experienced sensations. The conclusions are formulated as the following five dogmas:
1. To understand nervous function one needs to look at interactions at a cellular level, rather than either a more macroscopic or microscopic level, because behaviour depends upon the organized pattern of these intercellular interactions.

2. The sensory system is organized to achieve as complete a representation of the sensory stimulus as possible with the minimum number of active neurons.

neurons, each of which corresponds to a pattern of external events of the order of complexity of the events symbolized by a word.

5. High impulse frequency in such neurons corresponds to high certainty that the trigger feature is present.

The development of the concepts leading up to these speculative dogmas, their experimental basis, and some of their limitations are discussed.
V1 is highly overcomplete

The homunculus also has to face the problem that the image is often moving continuously, but is only represented by impulses at discrete moments in time. In these days he often has to deal with visual images derived from cinema screens and television sets that represent scenes sampled at quite long intervals, and we know that he does a good job at interpreting them even when the sample rate is only 16 s⁻¹, as in amateur movies. One only has to watch a kitten playing, a cat hunting, or a bird alighting at dusk among the branches of a tree to appreciate the importance and difficulty of the visual appreciation of motion. Considering this overwhelming importance it is surprising to find how slow are the receptors and how long is the latency for the message in the optic nerve, and even more surprising to find how well the system works in spite of this slowness.

Recent psychophysical work has improved our understanding of these problems. At one time it was thought that image motion aided resolution (Narshall & Talbot 1942), but this was hard to believe because of the blurring effect of the eye's long LGN afferents.
Sparse coding image model

(Olshausen & Field, 1996; Chen, Donoho & Saunders 1995)

\[ I(\vec{x}) = \sum_{i=1}^{M} a_i \phi_i(\vec{x}) + \epsilon(\vec{x}) \]
Energy function

\[ E = \frac{1}{2} |I - \Phi a|^2 + \lambda \sum_i C(a_i) \]

- preserve information
- be sparse
Locally Competitive Algorithm (LCA) minimizes the energy function
(Rozell, Johnson, Baraniuk & Olshausen, 2008)

\[ G_{ij} = \sum_{x} \phi_i(x) \phi_j(x) \]

\[ a_i = g(u_i) \]

\[ b_i = \sum_{x} \phi_i(x) I(x) \]

\[ \tau u_i + u_i = b_i - \sum_{j \neq i} G_{ij} a_j \]
Learned dictionary \( \{ \phi_i \} \)

Olshausen & Field (1996)
Sparse encoding of a time-varying image
Sparse coefficient activations form smooth trajectories in particular local subspaces
Basis functions tile the manifold of natural images in such a way that data points along the manifold are spanned by a small number of basis vectors.
Manifold flattening
Nonlinear Dimensionality Reduction by Locally Linear Embedding
Sam T. Roweis¹ and Lawrence K. Saul²

A Global Geometric Framework for Nonlinear Dimensionality Reduction
Joshua B. Tenenbaum,¹* Vin de Silva,² John C. Langford³

Science, 22 Dec. 2000
Local Linear Embedding (LLE)

1. Select neighbors
2. Reconstruct with linear weights
3. Map to embedded coordinates

\[ \varepsilon(W) = \sum_i \left| \tilde{X}_i - \sum_j W_{ij} \tilde{X}_j \right|^2 \]

\[ \Phi(Y) = \sum_i \left| \tilde{Y}_i - \sum_j W_{ij} \tilde{Y}_j \right|^2 \]
Local Linear Landmarks (LLL)
(Vladymyrov & Carreira-Perpinán, 2013)

\[ x = \Phi \alpha + n \quad \Rightarrow \quad y = P \alpha \]
From 1-sparse to k-sparse

1-sparse

Interpolation by KNN data vectors

k-sparse

Interpolation by landmarks or dictionary
‘Topographic ICA’
(Hyvärinen & Hoyer 2001)
Bubbles: a unifying framework for low-level statistical properties of natural image sequences

Aapo Hyvärinen, Jarmo Hurri, and Jaakko Väyrynen

Neural Networks Research Centre, Helsinki University of Technology, P.O. Box 9800, FIN-02015 HUT, Finland
Sparse Manifold Transform (Yubei Chen, Ph.D. thesis)

A) $x(\mathbb{R}^2)$

$k'$-sparse composition

D) $x_{local}(\mathbb{R}^2)$

locally 1-sparse component

B) sparse coding

$$\alpha = \arg \min_{\alpha} ||x - \Phi a||^2 + \lambda a$$

$$\alpha[\{1, \cdots, N\}]$$

$\Phi_1 \Phi_2 \Phi_3 \Phi_4$

manifold embedding

C) $\alpha_{TRUE}(M)$

$P_1 \cdots P_4$

dictionary local linear interpolation, or “steering”
Persistence
We seek a geometric mapping \( f : \Phi \rightarrow P \), s.t. each of the dictionary elements is mapped to a new vector, \( P_j = f(\Phi_j) \), where \( P_j \) is the \( j^{th} \) column of \( P \). Continuous temporal transformations in the input should have a linear flow on \( M \) and also in the geometrical embedding space.

\[
\Phi a \quad \rightarrow \quad P a
\]

We desire:
\[
Pa_t \approx \frac{1}{2}Pa_{t-1} + \frac{1}{2}Pa_{t+1}
\]

Objective function:
\[
\min_P \| PA D \|_F^2
\]

s.t. \( PV P^T = I \)

\[
V = \text{Cov}(a) \quad D = \begin{bmatrix}
1 & -\frac{1}{2} & 0 & 0 & \ldots & 0 \\
-\frac{1}{2} & 1 & -\frac{1}{2} & 0 & \ldots & 0 \\
0 & -\frac{1}{2} & 1 & -\frac{1}{2} & \ldots & 0 \\
0 & 0 & -\frac{1}{2} & 1 & \ldots & 0 \\
\vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\
0 & \ldots & 0 & 0 & \ldots & -\frac{1}{2} & 1
\end{bmatrix}
\]
The sparse manifold transform

\[ \hat{\alpha} = \arg \min_\alpha \| \beta - P \alpha \|^2 + \sum_i \lambda_i a_i \]

\[ \beta = P \alpha \]

\[ \hat{x} = \Phi \hat{\alpha} \]

\[ \alpha = \arg \min_\alpha \| x - \Phi \alpha \|^2 + \lambda \alpha \]
Simple example
To illustrate the distinction between the classical view of the activity of the representative pooling unit. The y-axis range indicates the total range of all pooling units. These outputs are sparse and tend to vary smoothly to form linear trajectories in the output space. The approach is effective for visualizing the underlying geometry, but less effective when the manifold space is nonlinear. At each time point, we can use a nearest neighbor (KNN) solver to select 300 landmarks on this disk as a dictionary. At each landmark, we set the corresponding row of \( P \) as a co-ordinate of the underlying manifold. It's equivalent to view a \( k \)-sparse function with the manifold as its domain. This turns a non-linear transformation into a Euclidean space would linearly merge all \( k \)-sparse functions to a dense and tend to vary smoothly to form linear trajectories in the embedding space because a topologically equivalent embedding rows as coordinates is equivalent to computing dot products of the pooling weight matrix with the lower dimensional data visualization using the second and third rows of \( P \). For each landmark, we set the corresponding \( \beta \)-th \( k \)-sparse functions defined on the underlying 2-D unit disk manifold. Visualizing the two rows that were used in the classical embedding are linear polynomials on the unit-disk, where the second and third rows have higher frequencies. As shown in Figure 3, since the first three rows that were used in the classical embedding are linear polynomials on the unit disk. The velocity is a random constant 2D vector and the starting locations are also random. We use clustering to select 300 landmarks on this disk as a dictionary. At each landmark, we set the corresponding row of \( P \) as a co-ordinate of the underlying manifold, shown in Figure 3, which closely resembles the Zernike ramp functions and the later rows have higher frequencies.
Simple example

k-sparse function embedding and recovery

4-sparse function \( \rightarrow \beta \rightarrow \) recovery
Encoding of a natural video sequence
We can produce reconstructions and dictionary visualizations of two layer-3 units that are each constructed by unwhitening matrix. We can also use the inverse transform example, the inverse transform from fine interpolations by repeatedly using the inverse operator, interpolation in layer-3 leads to a transformation, or a "steering", the reconstructions are unwhitened for visualization. D) Linear reconstruction from reconstruction from 

\[ \Phi^{(3)} \]

\[ \alpha^{(3)} \]

\[ \beta^{(2)} \]

\[ \Phi^{(2)} \]

\[ P^{(2)} \]

\[ \alpha^{(2)} \]

\[ \beta^{(1)} \]

\[ \Phi^{(1)} \]

\[ P^{(1)} \]

\[ \alpha^{(1)} \]

\[ \text{whitened image, } x \]

\[ \text{original image} \]

The rows are unique examples. All of 3) More abstract dictionary elements start to emerge features tend to be more global when we move to higher learning models produce a hierarchical code from image the representation over time in Figure 6. We apply the model to videos of natural scenes and demonstrate the stability of a topological relationship among sparse coding dictionary elements from time-varying input data, which we illustrate a hierarchical extension of the model in Figure 6. Discussion

- Nearly all of the dictionary elements in layer-3 span
- 
- in higher layers, e.g. layer-2 units tend to be more curved
- 
- of group structure by constructing a linear embedding of
- 
- learn the structure directly from the data. We utilize tempo-
- 
- hyv
- 
- Jarrett et al.
- 
- Karklin & Lewicki
- 
- Paiton et al.
Stacked Sparse Manifold Transform

reconstruction from layers 1, 2, 3

'flattening' at layer 3

\[
\begin{align*}
\Phi_k^{(3)} & \quad \Phi_l^{(3)} \\
& \quad \frac{1}{2} \Phi_k^{(3)} + \frac{1}{2} \Phi_l^{(3)}
\end{align*}
\]
Some neurophysiological effort has focused on characterizing IT neuronal tolerance to identity-preserving transformations and with varying levels of preference for one or the other face, analogous to what is observed in single unit recording in IT neurons. The same representation. It is important to note that, in contrast to the V1 simulation, we do not yet know how to generate single unit responses like this. Below, the responses of an example single unit are shown in response to the two faces undergoing one axis of pose variation.

Each colored rectangle represents a visual area, for the most part following the names and definitions used by Chowdhury and DeAngelis (2008). Wiesel's observation that visual cortex complex cells represented in each visual area and becomes better and better at directly supporting object recognition.

Figure 3

Manifolds are shown for the two objects (red and blue) undergoing two-axis pose variation. Figure 1b versus 3c)

Progress has been slow. But it is quite unexpected that this could be extracted using a complementary but qualitatively different approach. First, the object untangling perspective leads to a realization that single IT unit responses functions of individual neurons (i.e. the non-linear operators on the visual image) would, if successful, lead to an understanding the format of representation. Second, it suggests the immediate goal of determining how well each ventral stream neuronal representation preserves transformations – rather than the continuing problem, because such effects can be measured without the need to untangle at each ventral stream stage. Fourth, it suggests that the change in representation is likely to be significant, though its effects on vision in general, might be more meaningful (e.g. the predicted degree of errors a clear focus on the causes of tangling – identity-preserving transformations – rather than the continuing need to untangle at each ventral stream stage). But what is this transformation? In Figure 1c, object manifolds corresponding to the two objects are hopelessly tangled together. Perhaps in other monkeys MT would be able to find different areas involved in the two tasks. But it is quite unexpected that this could be extracted using a complementary but qualitatively different approach. First, the object untangling perspective leads to a realization that single IT unit responses functions of individual neurons (i.e. the non-linear operators on the visual image) would, if successful, lead to an understanding the format of representation.
Main points

• **Sparse coding** provides a foundation for manifold learning.

• A geometrical embedding of the dictionary may be learned by exploiting **temporal persistence** of structure in the visual world.

• Manifold flattening may be accomplished in a progressive manner by **successive stages** of sparse coding (dimensionality **expansion**) and linear projection (dimensionality **collapse**) in the ventral stream of visual cortex.