Lecture 15
Sparse Coding

Jochen Braun

Otto-von-Guericke-Universität Magdeburg,
Cognitive Biology Group

Engineering Neuroscience / Computational Neuroscience II
SS 2020

Credits: UFLDL Tutorial on Sparse Coding (ufldl.stanford.edu)
Andrew Ng ECCV10 Tutorial

15. Sparse coding

Efficient coding *is a computational hypothesis about neural representations and efficient use of neural resources to representing natural stimuli.* **Sparseness** *is a statistical property of natural stimuli (images, sounds, etc). Typical stimuli are composed of few features / feature combinations. Natural stimuli are not mixtures of Gaussian causes, so PCA is not right approach.* **Sparse coding** *is a computational hypothesis about sensory cortex. Neurons represent 'sparse' features and feature combinations. A sparse representation is overcomplete and minimizes the number of active neurons (not the total number of neurons). Sparse coding is also a* **practical computational strategy** *for learning an efficient representation (and causal model!) of complex signals. By learning an efficient causal model, cortex also better represents physical origins of sensory signals.*

Overview

- Efficient coding
- Natural image statistics
- Sparseness of neural responses
- Sparse coding theory
- Probabilistic interpretation (advanced)
- Learning (advanced)
- Summary

1. Efficient coding

- Sensory information is represented by neural activity in a brain region (e.g., in visual cortex).
- Response properties of neurons in this region determine nature of representation.
- How useful are particular representations with respect to behavioural goals?
- How efficient are particular representations with respect to a given sensory input statistics?

Attneave (1954) and Barlow (1961) proposed that information theory can link environmental statistics and neural responses through the concept of "coding efficiency".

Efficiency of causal models

Learning a causal model involves matching predicted $p[\boldsymbol{u}; G]$ to observed $p[\boldsymbol{u}]$. *Discrepancy* may be measured by Kullback-Leibler divergence

$$D_{KL}\left(p[\boldsymbol{u}], p[\boldsymbol{u}; G]\right) = \int_{obs} d\boldsymbol{u}\, p[\boldsymbol{u}]\, \ln \frac{p[\boldsymbol{u}]}{p[\mathbf{u}; \mathbf{G}]} \approx -\left\langle \ln p[\mathbf{u}; \mathbf{G}] \right\rangle_{obs} + const$$

Shannon theorem states that generative model G maximizing *likelihood* (or minimizing *discrepancy*)

$$L(G) = \left\langle \ln p[\mathbf{u}; \mathbf{G}] \right\rangle_{obs}$$

provides the most *efficient* way of encoding observations.

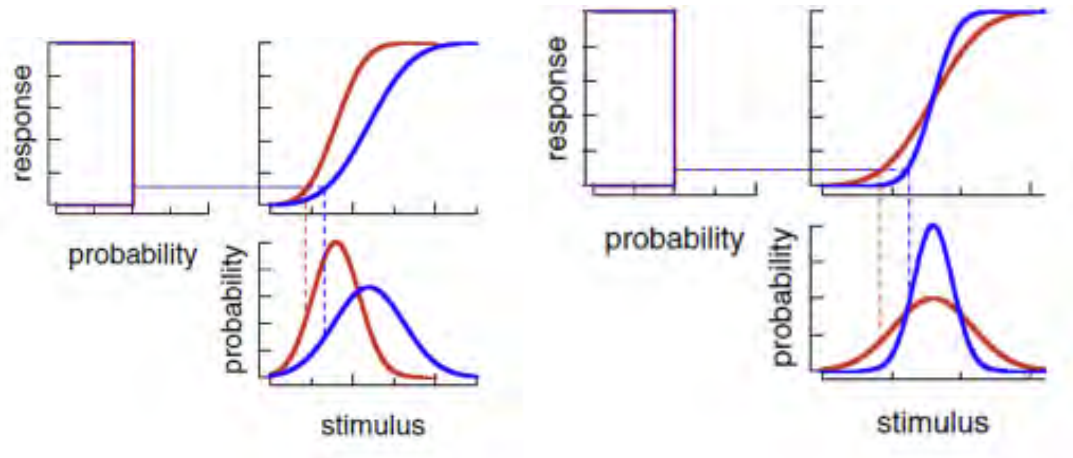
Same point will be made on slide 45.

Limitation of Gaussian models

Models $p[\mathbf{u}; \mathbf{G}]$ assuming Gaussian causes (principal component analysis, factor analysis) predict merely mean and variance of observations $p[\mathbf{u}]$.

Many sensory representations 'whiten' input signals, bringing mean to zero and variance to unity. For example, 'whitening' results from luminance and contrast adaptation.

From 'whitened' signals, Gaussian models cannot extract any information!



Gaussian input variance: principal components

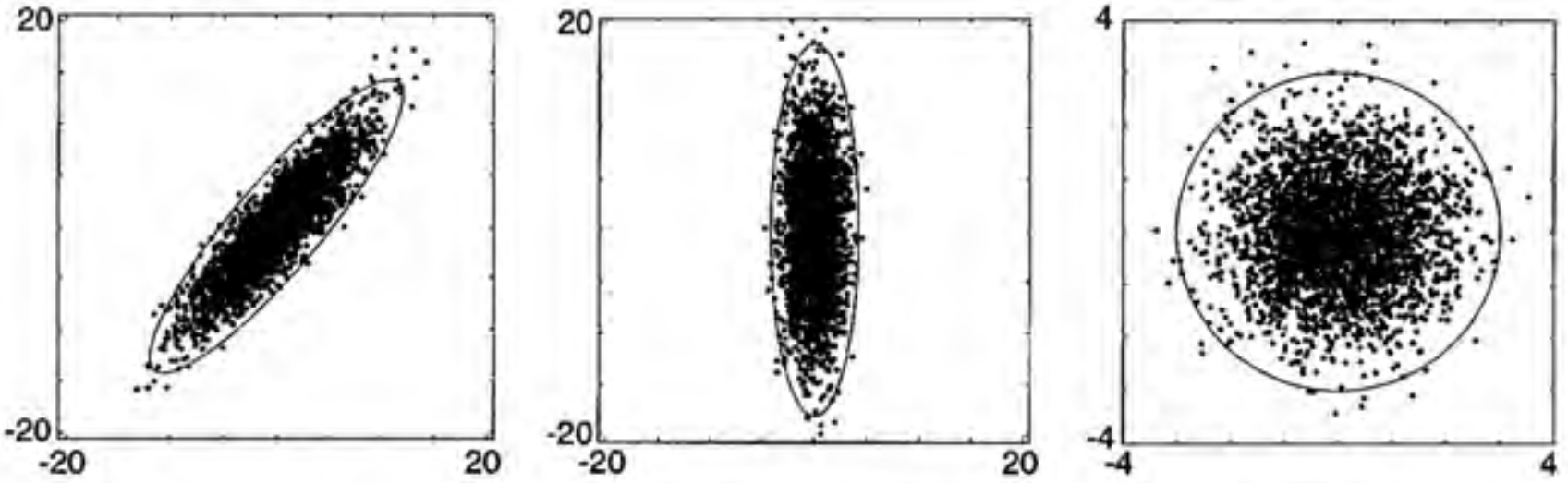


Figure 1: Illustration of principal component analysis on Gaussian-distributed data in two dimensions. (*a*) Original data. Each point corresponds to a sample of data drawn from the source distribution (i.e. a two-pixel image). The ellipse is three standard deviations from the mean in each direction. (*b*) Data rotated to principal component coordinate system. Note that the ellipse is now aligned with the axes of the space. (*c*) Whitened data. When the measurements are represented in this new coordinate system, their components are distributed as uncorrelated (and thus independent) univariate Gaussians.

Non-Gaussian variance: independent components

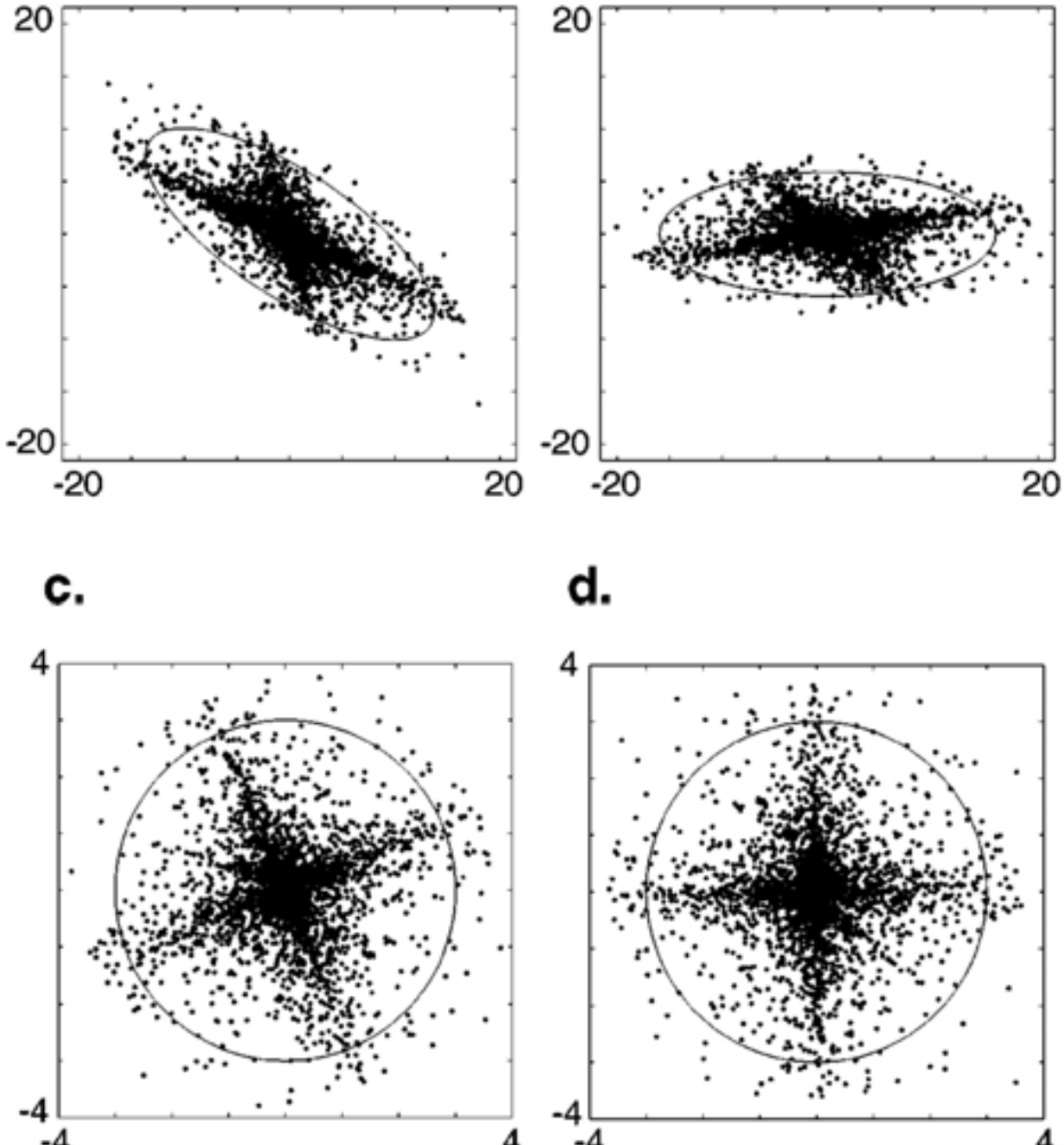


Figure 2 Illustration of principal component analysis and independent component analysis on non-Gaussian data in two dimensions. (*a*) Original data, a linear mixture of two non-Gaussian sources. As in Figure1, each point corresponds to a sample of data drawn from the source distribution, and the ellipse indicates three standard variations of the data in each direction. (*b*) Data rotated to principal component coordinate system. Note that the ellipse is now aligned with the axes of the space. (*c*) Whitened data. Note that the data are not aligned with the coordinate system. But the covariance ellipse is now a circle, indicating that the second-order statistics can give no further information about preferred axes of the data set. (*d*): Data after final rotation to independent component axes.

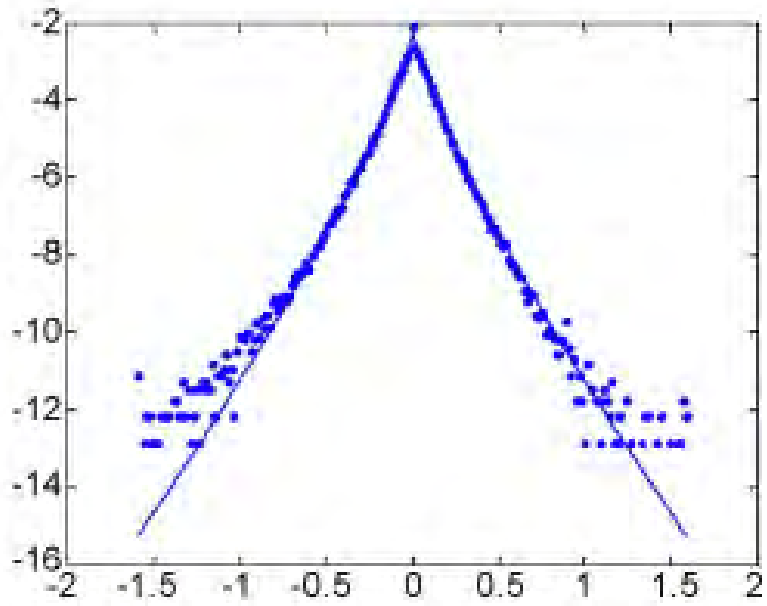
Points to note

- Neural representations may differ in usefulness and efficiency.
- Learning a causal model is equivalent to maximizing the *likelihood* that observations are generated by the model.
- A maximum-likelihood causal model provides the most efficient representation possible.
- Gaussian models are too simplistic for purposes of sensory representation.
- How can we go beyond Gaussian models?

2. Natural image statistics

Natural images (movies) contain contrast energy at different spatial (temporal) frequencies.

(a) natural image

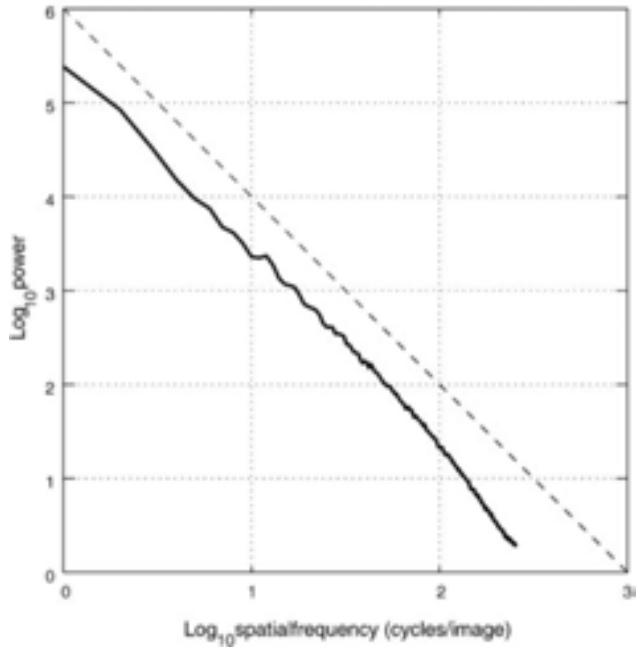


(b) *log* histogram

Mammalian retinal ganglion cells 'whiten' this power spectrum (Atick, Redlich, 1991).

Power law $1/f^2$

The spectral power of natural images falls with frequency, f, according to a power law, $1/f^2$ (Tolhurst, 1992; Ruderman & Bialek, 1994).



Power spectrum of a natural image (solid line) averaged over all orientations, compared with $1/f^2$ (dashed line).

Natural images

Natural images are *not* superpositions of independent Gaussian causes ...

Artificial images with Gaussian statistics do not look 'natural' and lack edges and structures. (Combine Gabor patterns with random location, orientation and frequency, Gaussian white noise, consistent with $1/f^2$ power law).

Whitened natural images do not look 'uniform' and retain the edges and structures of original image. (Decompose with Gabor filters, compensate for $1/f^2$ power law, and reassemble image).

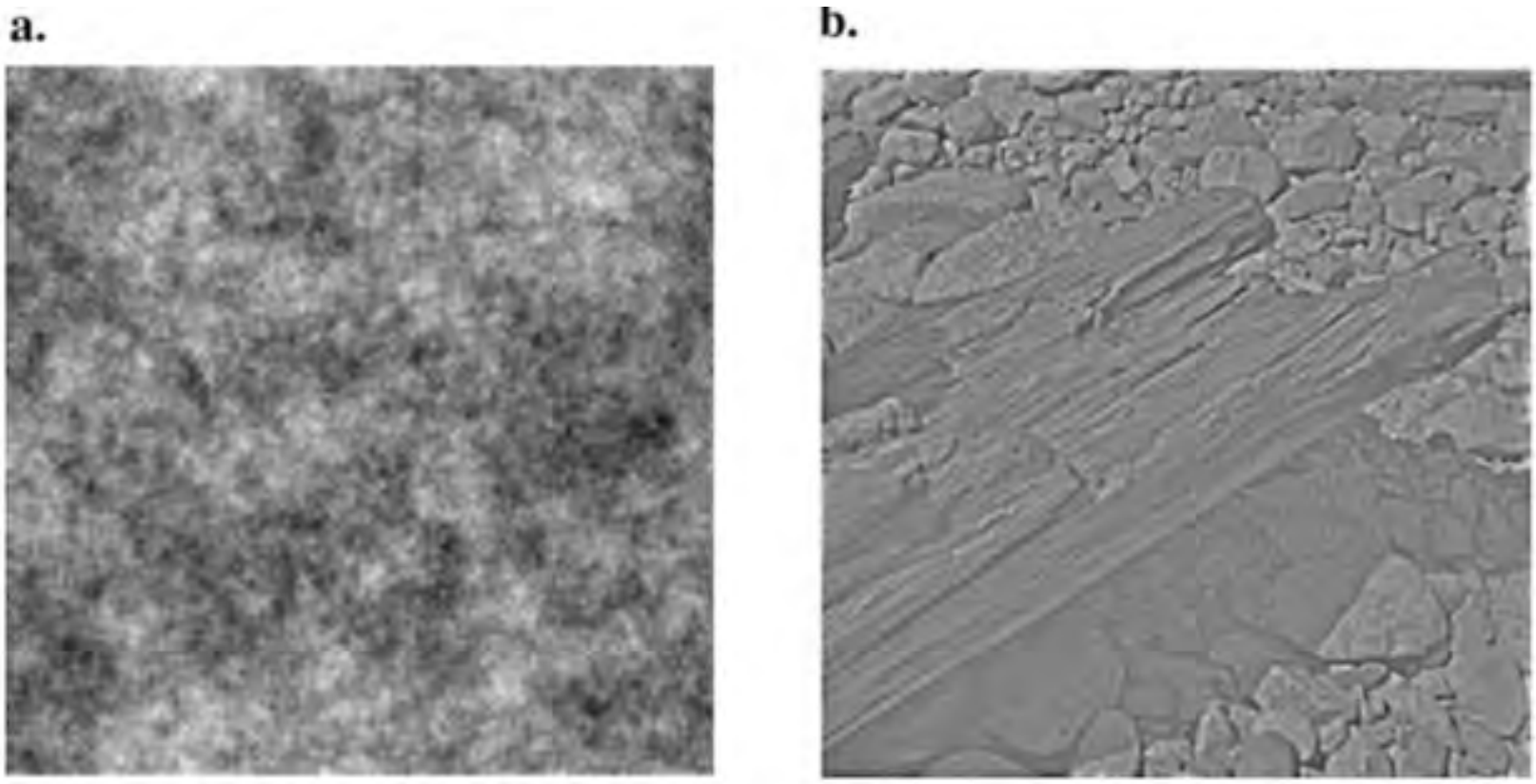


Figure 5 (*a*) Sample of 1/f Gaussian noise; (*b*) whitened natural image.

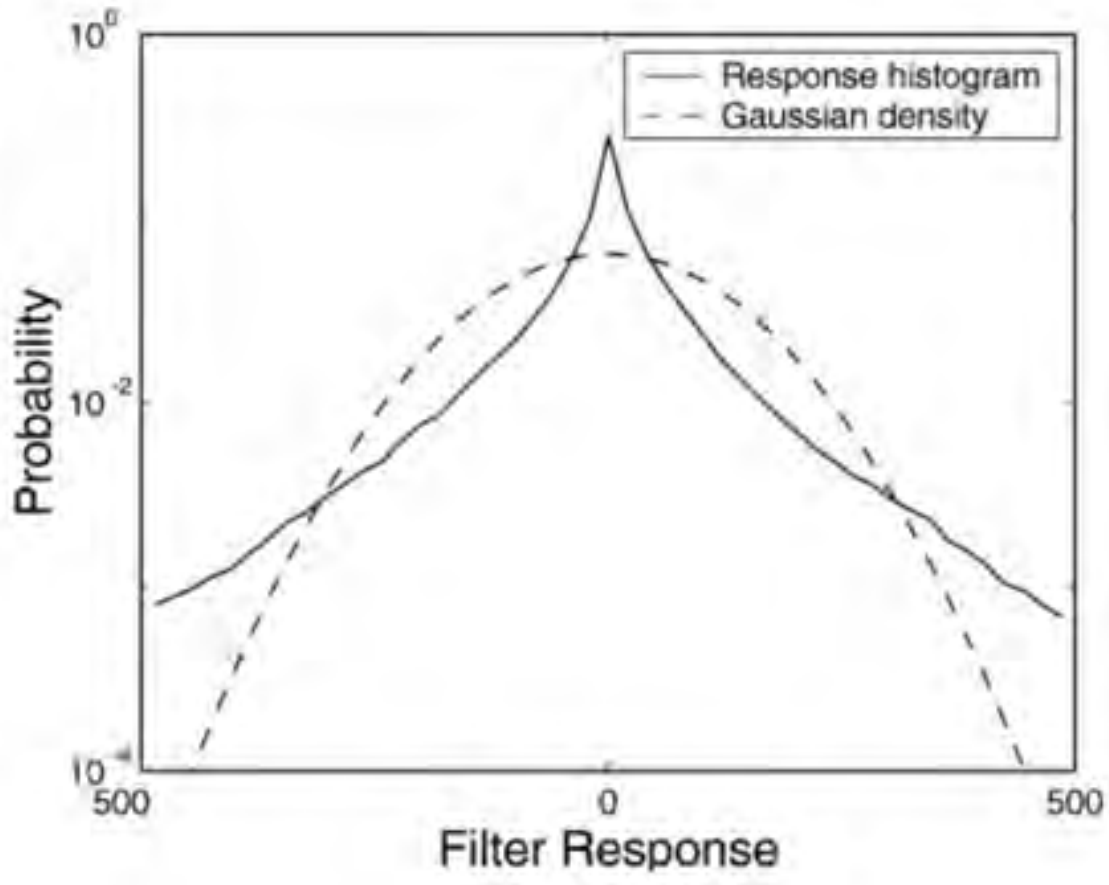
Gabor filter responses

Response distributions of oriented bandpass filters (e.g. Gabor filters) to natural images are characterized by 'sharp peaks' at zero and 'heavy tails' at large values (Field, 1987; Daugman, 1989).

Over a population of such filters, most responses are small, while only a few are large ("sparseness").

"Sparseness" of responses to natural images is maximal when filter tuning resembles that of visual cortical neurons (i.e. bandwidth of 0.5–1.5 octaves in spatial and temporal frequency).

Figure 6 Histogram of responses of a Gabor filter for a natural image, compared with a Gaussian distribution of the same variance.



Statistical dependencies of Gabor filter respones

Consider responses of different 'oriented bandpass filters' (e.g., different orientation and/or scale, but nearby visual locations) to natural images.

Typically, responses are not independent (Wegmann & Zetzche 1990, Simoncelli 1997, Simoncelli & Schwartz 1999).

Joint response distributions exhibit 'star' shape, so that large responses of one filter are associated with small response of the other, and *vice versa*.

Conditional response distributions exhibit 'bowtie' shape, so that large responses of one filter are associated with large variance of the other, and *vice versa*.

‘Star’ and ‘bowtie’

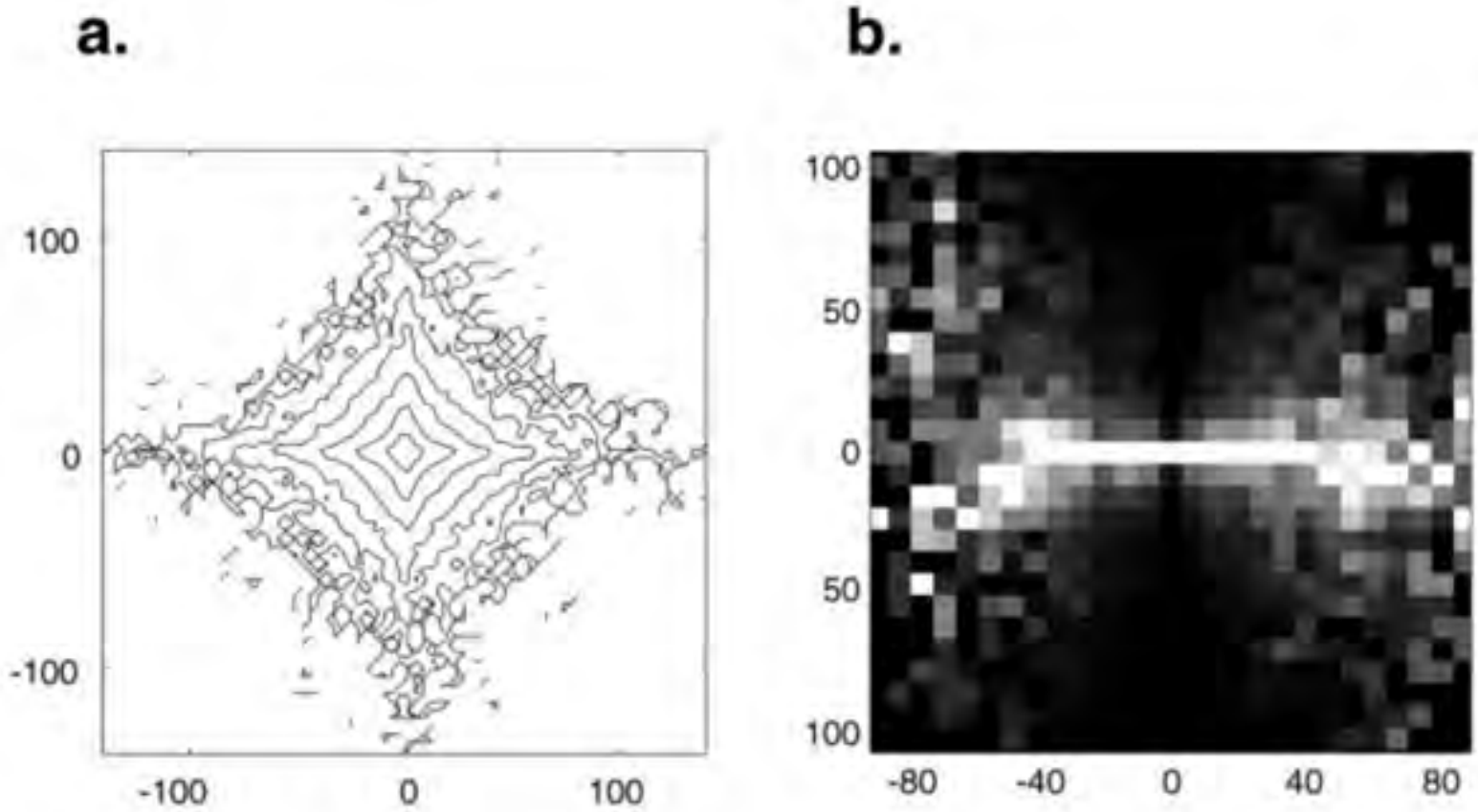


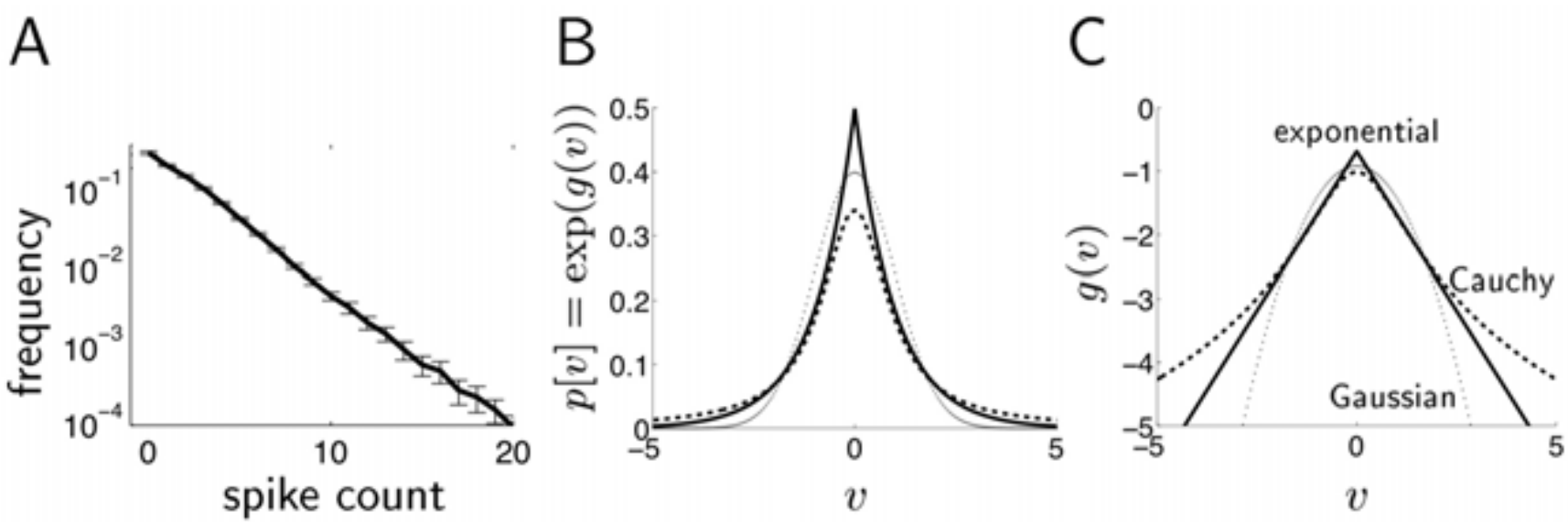
Figure 8 (*a*) Joint histogram of responses of two nonoverlapping receptive fields, depicted as a contour plot. (*b*) Conditional histogram of the same data. Brightness corresponds to probability, except that each column has been independently rescaled to fill the full range of display intensities (see Buccigrossi & Simoncelli 1999, Simoncelli & Schwartz 1999).

Summary

- Natural images carry information at all scales (frequencies f), but power decreases with frequency squared ($1/f^2$)
- Gabor filter responses to natural images are 'sparse'.
- Paired responses of different Gabor filters are 'doubly sparse' (mutually exclusive).
- Higher-order dependencies between filter pairs are informative about edges and structures.

3. Sparseness of neural responses

Response distribution of macaque IT neuron to natural videos is *exponential*.



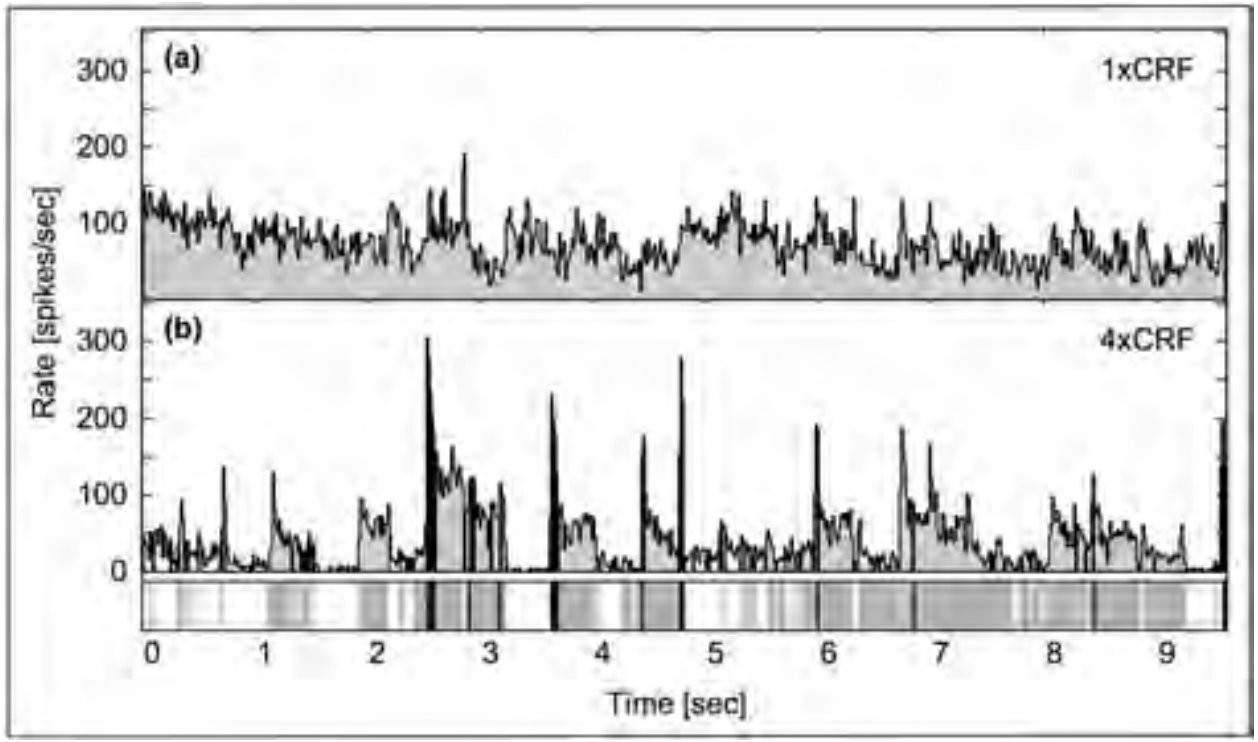
Exponential and Cauchy distributions are ‘sparse’ compared to Gaussian distribution. Here ‘sparseness’ means that responses are mostly near zero, but rarely far from zero.

Figure 10.4: Sparse distributions. A) Log frequency distribution of the activity of a macaque IT cell in response to a collection of three different videos. The size of the window used to calculate the spike counts was adjusted separately for each video so that, on average, there were two spikes per window. B) Three distributions $p[v]$ and $g[v] \equiv \ln p(v)$: exponential ($g(v) = -|v|$, solid); Cauchy ($g(v) = -\ln(1 + v^2)$, dashed); and Gaussian ($g(v) = -v^2/2$, dotted). C) The logarithms of the same three distributions. (A adapted from Baddeley et al, 1998.)

Visual cortex responses to natural movies

Example of a natural scene used as the source image for natural vision movies. White line represents simulated visual scan path. Image patches centered on the scan path were extracted to form the movie. Small white circle gives the classical receptive field (CRF) size; larger circle is four times the CRF diameter. **Vinje&Gallant, 2000**

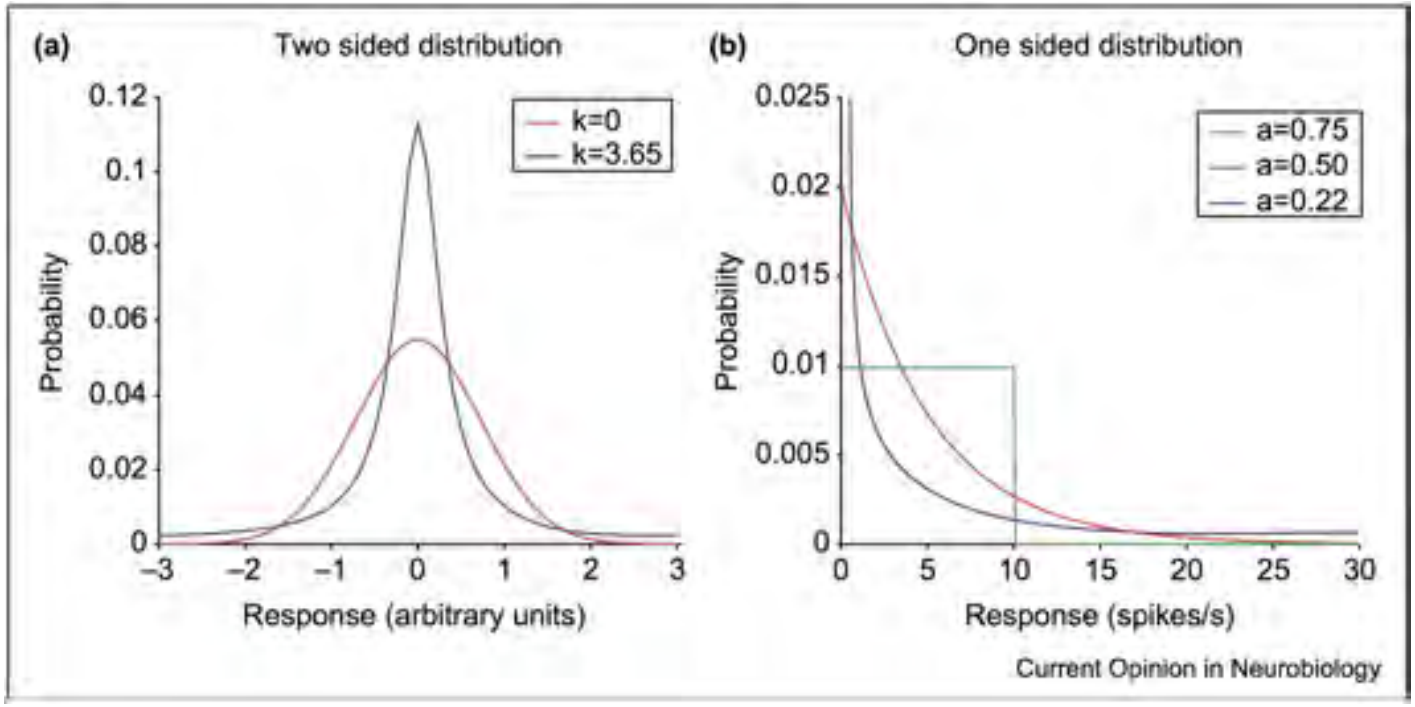
Responses in V1



Responses to natural scenes. Context in natural scenes sparsifies responses of V1 neurons. Shown is the average response of a neuron to multiple repetitions of a natural vision movie played just within the receptive field of the neuron (top) or the same movie but with additional spatial context extending into the receptive field surround (bottom). Context appears to make the neuron more selective to certain episodes within the movie sequence. **Vinje&Gallant, 2000**

Concept of sparseness

Neurons most likely fire few spikes, but very occasionally many spikes. The occasional strong response conveys substantial information. Such a response pattern is called ‘sparse’.
A ‘sparse’ response distribution is both ‘peaky’ and ‘heavy-tailed’.



Response distributions and sparseness. (a) Examples of two-sided response distributions for a unit that takes on both positive and negative values. A sparse representation would be consistent with a response distribution that is highly peaked at zero and with heavy tails (blue) compared to a Gaussian of the same variance (red). The former has positive kurtosis (k). (b) Examples of one-sided response distributions for a unit that takes on positive values only (e.g., firing rate). All distributions shown have the same mean firing rate. When plotted in this manner, a response distribution that is peaked at zero with heavy tails (blue) would be considered sparse, whereas a uniform response distribution (green) would not. The former has a low activity ratio (a) whereas the latter has a high activity ratio. The exponential distribution (red) lies somewhere in between. Note that measures of kurtosis (k) and activity ratio (a) are dimensionless. **Simoncelli & Olshausen, 2004**

Mean, variance, skewness, kurtosis

Mean and variance:

$$\bar{r} = \frac{1}{n}\sum_{i=1}^{n} r_i, \qquad \sigma^2 = \frac{1}{n}\sum_{i=1}^{n} (r_i - \bar{r})^2$$

Skewness and kurtosis (3^{rd} and 4^{th} central moments, normalized):

$$\gamma = \frac{1}{n}\sum_{i=1}^{n} \frac{(r_i - \bar{r})^3}{\sigma^3}, \qquad \kappa = \frac{1}{n}\sum_{i=1}^{n} \frac{(r_i - \bar{r})^4}{\sigma^4} \geq \gamma^2 + 1$$

For Gaussian distributions, skewness and kurtosis are zero. Kurtosis measures the frequency and size of 'outliers' ($r_i >> \sigma$) compared to 'inliers' ($r_i < \sigma$).

Typically, sparseness/heavy-tailed-ness is associated with values $\kappa > 3$.

One-sided-sparseness

A measure of (inverse) sparseness for one-sided distributions is (Rolls, Tovee, 1995)

$$a = \frac{\left(\frac{1}{n}\sum_{i=1}^{n} r_i\right)^2}{\frac{1}{n}\sum_{i=1}^{n} r_i^2}, \qquad sparse : \frac{1}{n} \qquad nonsparse : 1$$

For an extremely sparse distribution $\{1, 0, 0, 0, \ldots\}$, we have $a = {}^1/_n$. For an extremely non-sparse distribution $\{1, 1, 1, 1, \ldots\}$, we have $a = 1$.
A modified measure of sparseness is (Vinje, Gallant, 2001)

$$S = \frac{1 - a}{1 - {}^1/_n}, \qquad sparse : 1 \qquad nonsparse : 0$$

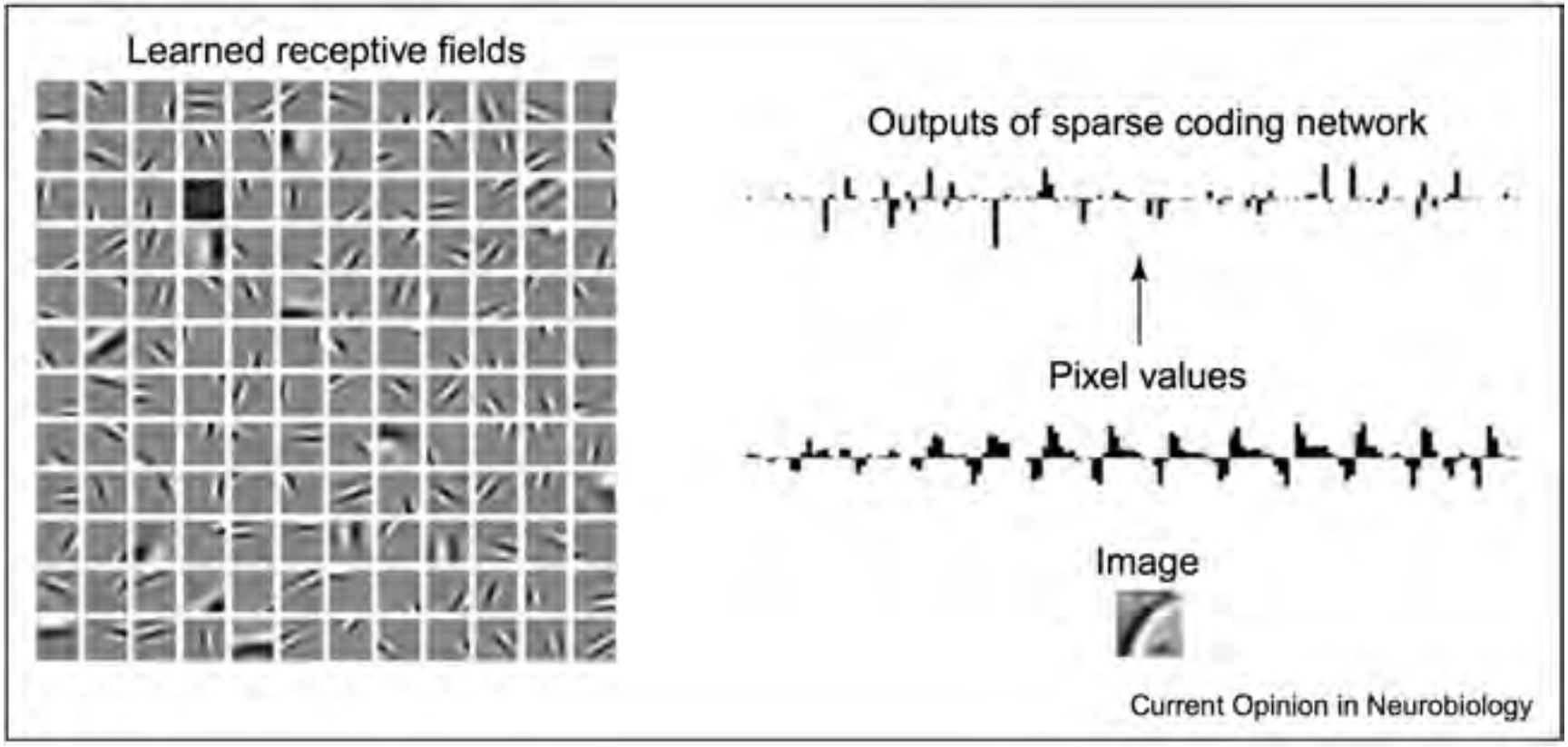
Size of representation

- Sparse representations minimize the number of *active* neurons, rather than the total number of neurons (which includes many *silent* neurons).
- This differs from Gaussian representations (principal component analysis), which minimize the number of represented causes.
- In fact, the size of cortical representations can be inflated. For example, for each cell in the visual thalamus there are approximately 40 cells in primary visual cortex.

Sparse coding model of simple cells

Simple-cell receptive fields may be related to sparse coding (Olshausen & Field, 1997).

- Seek an overcomplete set of linear basis-functions for natural images (i.e., set of receptive fields).
- Do not impose any particular shape on receptive fields (start with random pixel patterns).
- Seek to maximize sparsity of representation (number of zero coefficients).
- Seek to minimize representation error (preserving information).



- Train neural network on 500.000 image patches (12×12 pix).
- Basis functions resemble simple cell receptive fields.
- Spatially localized, oriented, and band-pass.
- Representation is sparser than original pixel representation!

Sparse coding of natural images. On the left is a set of receptive fields that are learnt by maximizing sparseness in the output of a neural network. Each patch shows the receptive field of a model neuron within a 12-by-12 pixel image patch. The network was trained on approximately half a million image patches (of the same size) extracted from whole images of natural scenes. The receptive fields that emerge from training are spatially localized, oriented, and bandpass (i.e., selective to spatial structure at a particular scale), similar to cortical simple cells. On the right is an example image patch and its encoding by the sparse coding network. The bar chart directly above the image patch shows the 144 pixel values contained in the patch. These input activities are transformed into a much sparser representation in the output of the network, shown in the bar chart at the top. The value of an output unit corresponds (roughly) to the degree of similarity between its receptive field and the input image. As the receptive fields are matched to the structures that typically occur in natural scenes, an image can usually be fully represented using a small number of active units. **Olshausen & Field, 2004**

4. Sparse coding theory

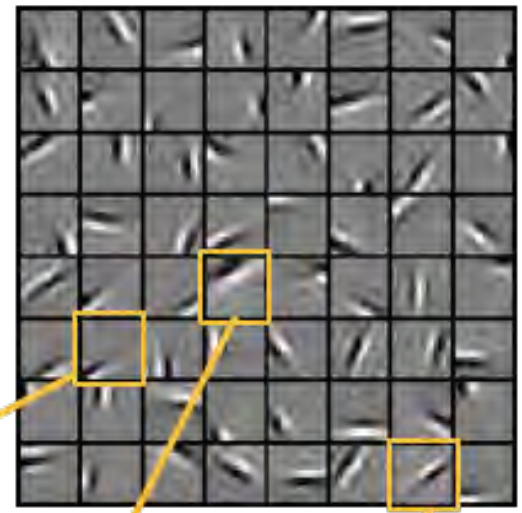
Sparse coding is a class of unsupervised methods for learning sets of over-complete bases to represent data efficiently. The aim of sparse coding is to find a set of basis vectors $\boldsymbol{\Phi}_i$ such that we can represent an n-dimenstional input vector $\boldsymbol{x} \in \mathcal{R}^n$ as a linear combination of these basis vectors:

$$\boldsymbol{x} = \sum_{i=1}^{k} a_i \, \boldsymbol{\Phi}_i$$

While techniques such as Principal Component Analysis (PCA) allow us to learn a **complete** set of basis vectors efficiently (i.e. such that $k = n$), we wish to learn an **over-complete set** of basis vectors to represent input vectors $\boldsymbol{x}$ (i.e. such that $k > n$).

Natural Images

Learned bases ($\phi_1, \ldots, \phi_{64}$)

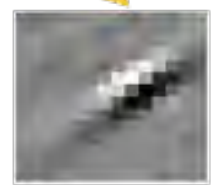


Test example

$\approx 0.8 *$

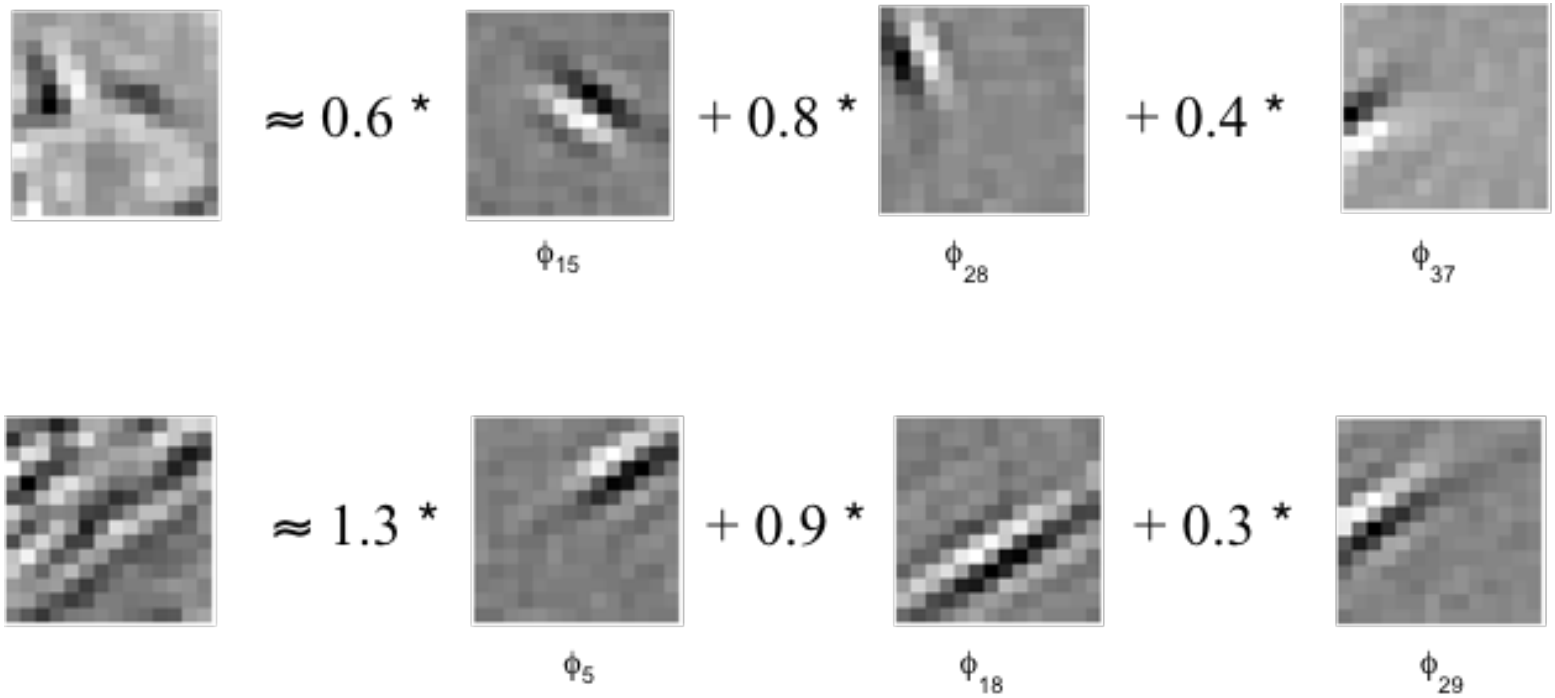
$+ 0.3 *$

$+ 0.5 *$



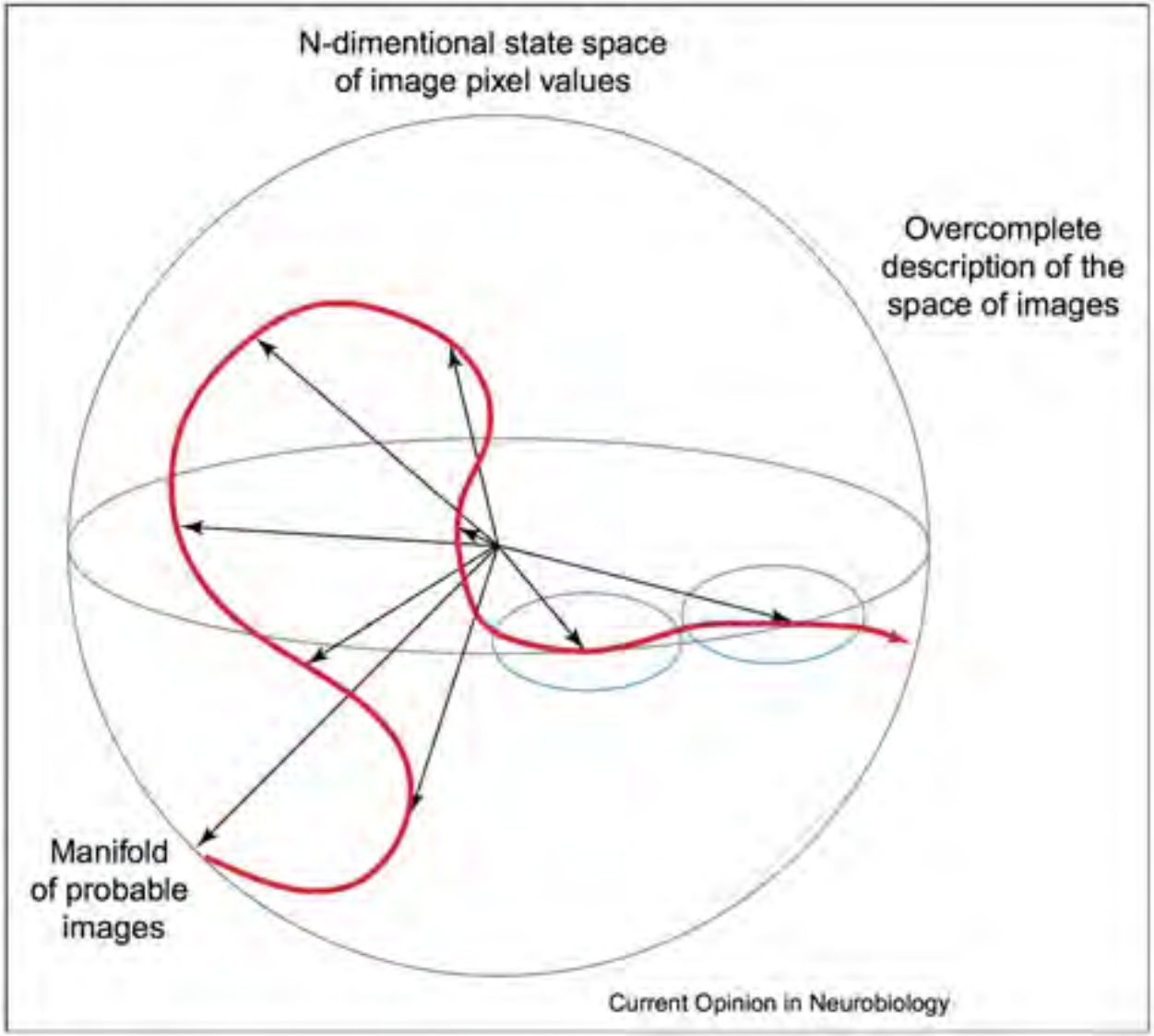
$x \approx 0.8 * \phi_{36} + 0.3 * \phi_{42} + 0.5 * \phi_{63}$

$[a_1, \ldots, a_{64}] = [0, 0, \ldots, 0, \mathbf{0.8}, 0, \ldots, 0, \mathbf{0.3}, 0, \ldots, 0, \mathbf{0.5}, \ldots]$



- Assume that edge-like patches are basic elements of a scene.
- Represent any scene as linear combination of edges.
- Obtain a more compact, higher-level representation than pixels.
- Sparse representation resembles visual receptive fields in primate cortex.

What's good about being over-complete?



State-space of natural scenes and over-complete codes. The sphere represents the N-dimensional state-space of natural scenes that is, the space of all possible images composed of N pixels. The axes of this space (not shown) are simply the pixel values of the image. Natural images are thought to lie along a low-dimensional manifold embedded in this space. The red curve represents the hypothetical trajectory of an image feature (such as an edge) as it would appear in this space as a result of translating over the pixel array. Each black arrow corresponds to the preferred feature of a neuron. The blue ellipses denote the response zone of the neuron that is, an image falling within this zone would cause the neuron to fire. The representation is overcomplete when there are more pattern vectors than input dimensions (image pixels). N pattern vectors would be sufficient to represent the manifold, but a sufficiently dense (highly overcomplete) tiling allows for a piecewise representation of a highly curved manifold, thus simplifying its representation for higher stages of analysis. **Olshausen & Field, 2004**

The advantage of having an over-complete basis is that our basis vectors are better able to capture structures and patterns inherent in the input data. However, with an over-complete basis, the coefficients a_i are no longer uniquely determined by the input vector $\boldsymbol{x}$.

We need **sparsity** as an additional criterion to resolve the degeneracy arising from over-completeness.

Here, we define sparsity as having few non-zero components or having few components not close to zero. The requirement that our coefficients a_i be sparse means that given a input vector, we would like as few of our coefficients to be far from zero as possible.

The choice of sparsity as a desired characteristic of our representation of the input data can be motivated by the observation that most sensory data such as natural images may be described as the superposition of a small number of atomic elements such as surfaces or edges.

An alternative justification is comparison to properties of the primary visual cortex and primary auditory cortex (see above).

Minimization function

Consider a set of m input vectors $\boldsymbol{x}^{(j)}$ (indexed by j) and an over-complete set of basis functions $\boldsymbol{\Phi}_i$ (indexed by i) and associated coefficients $a_i^{(j)}$ (indexed by both j and i)!

We seek a set of basis functions $\boldsymbol{\Phi}_i$ and coefficients $a_i^{(j)}$ that minimizes the sum of two cost functions:

$$\underbrace{\sum_{j=1}^{m} \left\| \boldsymbol{x}^{(j)} - \sum_{i=1}^{k} a_i^{(j)} \boldsymbol{\Phi}_i \right\|^2}_{\textit{quality}} + \lambda \underbrace{\sum_{i=1}^{k} S\left(a_i^{(j)}\right)}_{\textit{sparseness}}$$

where $S(\cdot)$ is a sparsity cost function which penalizes $a_i^{(j)}$ for being far from zero.

The first term favours quality of representation of $\boldsymbol{x}^{(j)}$ and the second term favours sparseness of representation. The constant λ determines the relative importance of both factors.

Sparsity penalty $S(\cdot)$

L^0-norm, number of non-zero elements (difficult to optimize)

$$\sum_i S(a_i) = \|\boldsymbol{a}\|_0 \qquad S(a_i) = \begin{cases} 1 & |a_i| > 0 \\ 0 & |a_i| = 0 \end{cases}$$

L^1-norm (easier):

$$\sum_i S(a_i) = \|\boldsymbol{a}\|_1 = |a_1| + |a_2| + \ldots + |a_k|$$

Log penalty (differentiable):

$$\sum_i S(a_i) = \sum_i \log\left(1 + a_i^2\right)$$

L^p-norm:

$$\|x\|_p = \left(|x_1|^p + |x_2|^p + \ldots + |x_n|^p\right)^{1/p}$$

Power of basis functions

Formally, the sparsity penalty can be avoided by scaling down a_i and scaling up $\mathbf{\Phi}_i$. To prevent this, the power of basis functions $\|\mathbf{\Phi}_i\|^2$ must be constrained to some ceiling C.
The complete function to be maximized is then

$$\min_{\boldsymbol{a}_i, \mathbf{\Phi}_i} \left[\sum_{j=1}^{m} \left\| \boldsymbol{x}^{(j)} - \sum_{i=1}^{k} a_i^{(j)} \mathbf{\Phi}_i \right\|^2 + \lambda \sum_{i=1}^{k} S\left(a_i^{(j)}\right) \right]$$

subject to

$$\|\mathbf{\Phi}_i\|^2 \leq C, \qquad \forall i$$

The minimization is performed alternatingly with respect to $\mathbf{\Phi}_i$ (easier) and with respect to $a_i^{(j)}$ (harder).

Feature sign search

Minimize with respect to $\mathbf{a} = (a_1, a_2, a_3)$ (simplified example):

$$\min_{\boldsymbol{a}} \left[\|\boldsymbol{a} - (a_1\boldsymbol{\Phi}_1 + a_2\boldsymbol{\Phi}_2 + a_3\boldsymbol{\Phi}_3)\| + \lambda\left(|a_1| + |a_2| + |a_3|\right) \right]$$

Suppose you know the signs of the a_i: $a_1 \geq 0, a_2 \leq 0, a_2 \geq 0$. This simplies the problem to

$$\min_{\boldsymbol{a}} \left[\|\boldsymbol{a} - (a_1\boldsymbol{\Phi}_1 + a_2\boldsymbol{\Phi}_2 + a_3\boldsymbol{\Phi}_3)\| + \lambda\left(a_1 - a_2 + a_3\right) \right]$$

This quadratic equation in $\boldsymbol{a}$ can be solved efficiently in closed form.

Algorithm: guess signs $(+, 0, -)$ of a_i, solve in closed form, refine guesses for signs, ...

Summary

- Sparse coding is a computational method for representing complex observations.
- Observations are encoded as linear combinations of basis vectors.
- Basis vectors are more numerous than necessary (over-complete).
- Chosen to minimize the number of terms in linear combinations (sparseness).

5. Causal model interpretation (advanced)

We have considered sparse coding as a computational method for describing complex observations. But might there not be some truth in a sparse description? Might it not reveal underlying physical causes?

Consider a sparse coding scheme as a causal model (e.g., for observable images):

$$P_{gen}(\boldsymbol{x}|\boldsymbol{\Phi}), \qquad \boldsymbol{x} = \sum_{i=1}^{k} a_i \boldsymbol{\Phi}_i + \nu(\boldsymbol{x})$$

Model natural images $\mathbf{x}$ as a linear superposition of k independent source features $\boldsymbol{\Phi}_i$ with sparse coefficients $\boldsymbol{a}$ and some additive noise ν!

Kullback-Leibler divergence of distributions

Our goal is to find a set of basis feature vectors $\mathbf{\Phi}$ such that the distribution of images $P_{gen}(\boldsymbol{x}|\mathbf{\Phi})$ is as close as possible to the empirical distribution of our input data $P_{obs}(\boldsymbol{x})$.

One method is to minimize the KL divergence between P_{obs} and $P_{gen}(\boldsymbol{x}|\mathbf{\Phi})$:

$$D\left[P_{obs}(\boldsymbol{x}) \parallel P(\boldsymbol{x}|\mathbf{\Phi})\right] = \int P_{obs}(\boldsymbol{x}) \ln \left[\frac{P_{obs}(\boldsymbol{x})}{P_{gen}(\boldsymbol{x}|\mathbf{\Phi})}\right] d\boldsymbol{x}$$

Since the empirical distribution $P_{obs}(\boldsymbol{x})$ is constant across our choice of $\mathbf{\Phi}$, this is equivalent to maximizing the log-likelihood of $P_{gen}(\boldsymbol{x}|\mathbf{\Phi})$.

Same point was made on slide 5.

Gaussian noise & independence

Assuming Gaussian white noise ν with variance σ^2, we have that

$$P(\boldsymbol{x}|a, \boldsymbol{\Phi}) = \frac{1}{Z} \exp\left[-\frac{\left(\boldsymbol{x} - \sum_{i=1}^{k} a_i \boldsymbol{\Phi}_i\right)^2}{2\sigma^2}\right]$$

In order to determine the distribution $P_{gen}(\boldsymbol{x}|\boldsymbol{\Phi})$, we also need to specify the prior distribution $P(\boldsymbol{a})$. Assuming independence of source features, we can factorize the prior probability as

$$P(\boldsymbol{a}) = \prod_{i=1}^{k} P(a_i)$$

Sparsity constraint

At this point, we introduce sparsity in our generative model, i.e., the assumption that any single image is likely to be the product of relatively few source features. Thus, we would like the distribution of a_i to peak at zero and have high kurtosis.

A convenient parameterization for this is

$$P(a_i) = \frac{1}{Z} \exp\left(-\beta S(a_i)\right)$$

Where $S(a_i)$ is a function determining the shape of the prior distribution.

Maximization function

Having defined $P(\boldsymbol{x}|a, \boldsymbol{\Phi})$ and $P(\boldsymbol{a})$, we can write the probability of the data $\boldsymbol{x}$ under the generative model defined by $\boldsymbol{\Phi}$ as

$$P_{gen}(\boldsymbol{x}|\boldsymbol{\Phi}) = \int P(\boldsymbol{x}|a, \boldsymbol{\Phi})\, P(\boldsymbol{a})\, da$$

and our problem reduces to finding

$$\boldsymbol{\Phi}^{\star} = \max_{\boldsymbol{\Phi}} E\left[\log\left(P_{gen}(\boldsymbol{x}|\boldsymbol{\Phi})\right)\right]_{\boldsymbol{x}}$$

where $E[\cdot]_{\boldsymbol{x}}$ denotes expectation over our input observations.

Practicalities

Unfortunately, the integral over $\boldsymbol{a}$ to obtain $P_{gen}(\boldsymbol{x}|\boldsymbol{\Phi})$ is generally intractable. We note though that if the distribution of $P_{gen}(\boldsymbol{x}|\boldsymbol{\Phi})$is sufficiently peaked (w.r.t. $\boldsymbol{a}$), we can approximate its integral with the maximum value of $P_{gen}(\boldsymbol{x}|\boldsymbol{\Phi})$ and obtain a approximate solution

$$\boldsymbol{\Phi}^{\star} \approx \max_{\boldsymbol{\Phi}} E\left[\max_{\boldsymbol{a}} \log\left(P_{gen}(\boldsymbol{x}|\boldsymbol{\Phi})\right)\right]_{\boldsymbol{x}}$$

As before, we need to constrain the power of $\boldsymbol{\Phi}$ to prevent scaling down a_i and scaling up $\boldsymbol{\Phi}_i$.

Compare original theory

Finally, we can recover the sparse coding cost function (of slide 41) by defining the energy function of this linear generative model as

$$\begin{aligned} E(\boldsymbol{x}, \boldsymbol{a}|\boldsymbol{\Phi}) &= -\log\left[P_{gen}(\boldsymbol{x}|\boldsymbol{\Phi}, \boldsymbol{a})\, P(\boldsymbol{a})\right] \\ &= \sum_{j=1}^{m} \left\| \boldsymbol{x}^{(j)} - \sum_{i=1}^{k} a_i^{(j)}\, \boldsymbol{\Phi}_i \right\|^2 + \lambda \sum_{i=1}^{k} S\left(a_i^{(j)}\right) \end{aligned}$$

where $\lambda = 2\,\sigma^2\,\beta$ and irrelevant constants have been hidden.

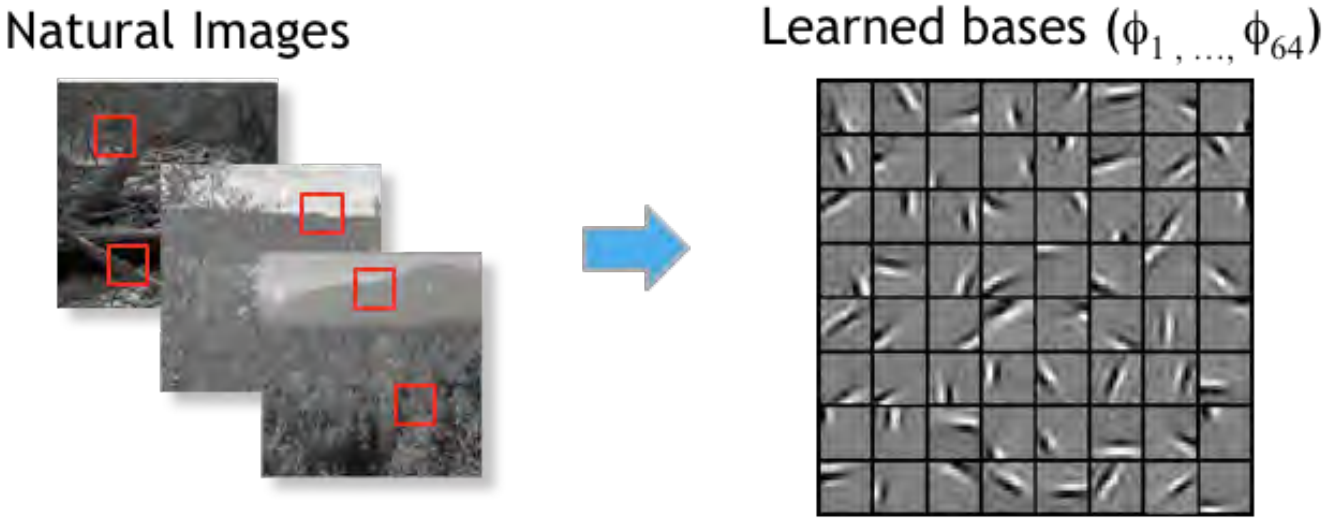
Since maximizing the log-likelihood is equivalent to minimizing the energy function, we recover the original optimization problem:

$$\boldsymbol{\Phi}^\star, \boldsymbol{a}^\star = \min_{\boldsymbol{\Phi}, \boldsymbol{a}} \left[\sum_{j=1}^{m} \left\| \boldsymbol{x}^{(j)} - \sum_{i=1}^{k} a_i^{(j)} \boldsymbol{\Phi}_i \right\|^2 + \lambda \sum_{i=1}^{k} S\left(a_i^{(j)}\right) \right]$$

Using a probabilistic approach, it can also be seen that the choices of the L^1 penalty $|a_i|$ and the log penalty $\log(1 + a_i^2)$ for $S(\cdot)$ correspond to the use of the Laplacian $P(a_i) \propto \exp(-\beta|a_i|)$ and the Cauchy prior $P(a_i) \propto \beta/|1 + a_i^2|$, respectively.

Thus, sparse coding schemes provide efficient causal models and may capture some truth about the physical origin of our observations!

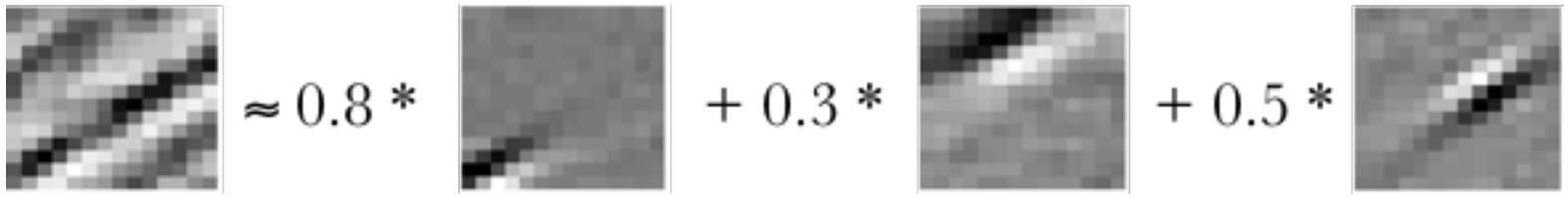
Summary: Training phase



- Input: Images $\boldsymbol{x}^{(1)}, \boldsymbol{x}^{(2)}, \ldots, \boldsymbol{x}^{(m)}$ in $\mathcal{R}^{n \times n}$
- Learn: Dictionary of basis functions $\boldsymbol{\Phi}^{(1)}, \boldsymbol{\Phi}^{(2)}, \ldots, \boldsymbol{\Phi}^{(k)}$ in $\mathcal{R}^{n \times n}$

$$\min_{\boldsymbol{\Phi}, \boldsymbol{a}} \left[\sum_{j=1}^{m} \left\| \boldsymbol{x}^{(j)} - \sum_{i=1}^{k} a_i^{(j)} \boldsymbol{\Phi}_i \right\|^2 + \lambda \sum_{i=1}^{k} S\left(a_i^{(j)}\right) \right]$$

Summary: Recognition phase



- Input: Novel images $\boldsymbol{x}$ and previously learned $\boldsymbol{\Phi}_i$.
- Output: Sparse representation $[a_1, a_2, \ldots, a_k]$ of image $\boldsymbol{x}$:

$$\min_{\boldsymbol{a}} \left[\left\| \boldsymbol{x} - \sum_{i=1}^{k} a_i \, \boldsymbol{\Phi}_i \right\|^2 + \lambda \sum_{i=1}^{k} S\left(a_i\right) \right]$$

6. Learning (advanced)

Learning a set of basis vectors $\mathbf{\Phi}$ using sparse coding consists of performing two separate optimizations, the first being an optimization over coefficients a_i for each training example $\boldsymbol{x}$ and the second an optimization over basis vectors $\mathbf{\Phi}$ across many training examples at once.

Assuming an L^1 sparsity penalty, learning $a_i^{(j)}$ reduces to solving a L^1 regularized least squares problem which is convex in $a_i^{(j)}$ for which several techniques have been developed (convex optimization software such as CVX can also be used to perform L^1 regularized least squares).

Assuming a differentiable $S(\cdot)$ such as the log penalty, gradient-based methods such as conjugate gradient methods can also be used.

Learning a set of basis vectors with a L^2 norm constraint also reduces to a least squares problem with quadratic constraints which is convex in $\mathbf{\Phi}$. Standard convex optimization software (e.g. CVX) or other iterative methods can be used to solve for $\mathbf{\Phi}$, although significantly more efficient methods such as solving the Lagrange dual have also been developed.

As described above, a significant limitation of sparse coding is that even after a set of basis vectors have been learnt, in order to 'encode' a new data example, optimization must be performed to obtain the required coefficients. This significant 'runtime' cost means that sparse coding is computationally expensive to implement even at test time, especially compared to typical feedforward architectures.

Overall Summary

- **'Efficient coding' is a computational hypothesis about neural representations and efficient use of neural resources to representing natural stimuli.**
- **'Sparseness' appears to be a property of natural stimulus (images, sounds, etc) statistics. Not all feature combinations occur equally often. A subset of combinations occurs with comparable frequency, other combinations almost never.**
- **'Sparse coding' is a computational hypothesis about visual (and auditory) cortex. Sensory neurons efficiently represent 'sparse' features and feature combinations.**
- **'Sparse coding' is also practical computational strategy for efficient stimulus representation.**

Next & last:

Independent Component Analysis