LETTER Communicated by Odelia Schwartz

On the Analysis and Interpretation of Inhomogeneous Quadratic Forms as Receptive Fields

Pietro Berkes
p.berkes@biologie.hu-berlin.de
Laurenz Wiskott
l.wiskott@biologie.hu-berlin.de
Institute for Theoretical Biology, Humboldt University Berlin, D-10115 Berlin, Germany

In this letter, we introduce some mathematical and numerical tools to analyze and interpret inhomogeneous quadratic forms. The resulting characterization is in some aspects similar to that given by experimental studies of cortical cells, making it particularly suitable for application to second-order approximations and theoretical models of physiological receptive fields. We first discuss two ways of analyzing a quadratic form by visualizing the coefficients of its quadratic and linear term directly and by considering the eigenvectors of its quadratic term. We then present an algorithm to compute the optimal excitatory and inhibitory stimuli—those that maximize and minimize the considered quadratic form, respectively, given a fixed energy constraint. The analysis of the optimal stimuli is completed by considering their invariances, which are the transformations to which the quadratic form is most insensitive, and by introducing a test to determine which of these are statistically significant. Next we propose a way to measure the relative contribution of the quadratic and linear term to the total output of the quadratic form. Furthermore, we derive simpler versions of the above techniques in the special case of a quadratic form without linear term. In the final part of the letter, we show that for each quadratic form, it is possible to build an equivalent two-layer neural network, which is compatible with (but more general than) related networks used in some recent articles and with the energy model of complex cells. We show that the neural network is unique only up to an arbitrary orthogonal transformation of the excitatory and inhibitory subunits in the first layer.

1 Introduction

Recent research in neuroscience has seen an increasing number of extensions of established linear techniques to their nonlinear equivalent in both experimental and theoretical studies. This is the case, for example, for spatiotemporal receptive field estimates in physiological studies (see

Neural Computation **18, 1868–1895** (2006)

Simoncelli, Pillow, Paninski, & Schwartz, 2004, for a review) and information-theoretical models like principal component analysis (PCA) (Schölkopf, Smola, & Müller, 1998) and independent component analysis (ICA) (see Jutten & Karhunen, 2003, for a review). Additionally, new nonlinear unsupervised algorithms have been introduced, for example, slow feature analysis (SFA) (Wiskott & Sejnowski, 2002). The study of the resulting nonlinear functions can be a difficult task because of the lack of appropriate tools to characterize them qualitatively and quantitatively.

During a recent project concerning the self-organization of complex cell receptive fields in the primary visual cortex (V1) (Berkes & Wiskott, 2002, 2005b; see section 2), we developed some of these tools to analyze quadratic functions in a high-dimensional space. Because of the complexity of the methods, we describe them here in a separate letter. The resulting characterization is in some aspects similar to that given by physiological studies, making it particularly suitable to be applied to the analysis of nonlinear receptive fields.

We are going to focus on the analysis of the inhomogeneous quadratic form

$$g(\mathbf{x}) = \frac{1}{2}\mathbf{x}^T\mathbf{H}\mathbf{x} + \mathbf{f}^T\mathbf{x} + c\,, \tag{1.1}$$

where $\mathbf{x}$ is an N-dimensional input vector, $\mathbf{H}$ an $N \times N$ matrix, $\mathbf{f}$ an N-dimensional vector, and c a constant. Although some of the mathematical details of this study are specific to quadratic forms only, it should be straightforward to extend most of the methods to other nonlinear functions while preserving the same interpretations. In other contexts, it might be more useful to approximate the function under consideration by a quadratic form using a Taylor expansion up to the second order and then apply the algorithms described here.

In experimental studies, quadratic forms occur naturally as a second-order approximation of the receptive field of a neuron in a Wiener expansion (Marmarelis & Marmarelis, 1978; van Steveninck & Bialek, 1988; Lewis, Henry, & Yamada, 2002; Schwartz, Chichilnisky, & Simoncelli, 2002; Touryan, Lau, & Dan, 2002; Rust, Schwartz, Movshon, & Simoncelli, 2004; Simoncelli et al., 2004). Quadratic forms were also used in various theoretical articles, either explicitly (Hashimoto, 2003; Bartsch & Obermayer, 2003) or implicitly in the form of neural networks (Hyvärinen & Hoyer, 2000, 2001; Körding, Kayser, Einhäuser, & König, 2004). The analysis methods used in these studies are discussed in section 10.

Table 1 lists some important terms and variables used throughout the article. We will refer to $\frac{1}{2}\,\mathbf{x}^T\mathbf{H}\mathbf{x}$ as the *quadratic term*, to $\mathbf{f}^T\mathbf{x}$ as the *linear term*, and to c as the *constant term* of the quadratic form. Without loss of generality, we assume that $\mathbf{H}$ is a symmetric matrix, since if necessary we can substitute

Table 1: Definitions of Some Important Terms.

N	Number of dimensions of the input space
$\langle\cdot\rangle_t$	Mean over time of the expression between the two brackets
$\mathbf{x}$	Input vector
$g, \tilde{g}$	The considered inhomogeneous quadratic form and its restriction to a sphere
$\mathbf{H}, \mathbf{h}_i$	$N \times N$ matrix of the quadratic term of the inhomogeneous quadratic form (see equation 1.1) and ith row of $\mathbf{H}$ (i.e., $\mathbf{H} = (\mathbf{h}_1, \ldots, \mathbf{h}_N)^T$). $\mathbf{H}$ is assumed to be symmetric.
$\mathbf{v}_i, \mu_i$	ith eigenvector and eigenvalue of $\mathbf{H}$, sorted by decreasing eigenvalues (i.e., $\mu_1 \geq \mu_2 \geq \ldots \geq \mu_N$)
$\mathbf{V}, \mathbf{D}$	The matrix of the eigenvectors $\mathbf{V} = (\mathbf{v}_1, \ldots, \mathbf{v}_N)$ and the diagonal matrix of the eigenvalues, so that $\mathbf{V}^T\mathbf{HV} = \mathbf{D}$
$\mathbf{f}$	N-dimensional vector of the linear term of the inhomogeneous quadratic form (see equation 1.1)
c	Scalar value of the constant term of the inhomogeneous quadratic form (see equation 1.1)
$\mathbf{x}^+, \mathbf{x}^-$	Optimal excitatory and inhibitory stimuli, $\Vert\mathbf{x}^+\Vert = \Vert\mathbf{x}^-\Vert = r$

$\mathbf{H}$ in equation 1.1 by the symmetric matrix $\frac{1}{2}$ $(\mathbf{H} + \mathbf{H}^\mathbf{T})$ without changing the values of the function g. We define $\mu_1, \ldots, \mu_N$ to be the eigenvalues to the eigenvectors $\mathbf{v}_1, \ldots, \mathbf{v}_N$ of $\mathbf{H}$ sorted in decreasing order $\mu_1 \geq \mu_2 \geq \ldots \geq \mu_N$. $\mathbf{V} = (\mathbf{v}_1, \ldots, \mathbf{v}_N)$ denotes the matrix of the eigenvectors and $\mathbf{D}$ the diagonal matrix of the corresponding eigenvalues, so that $\mathbf{V}^T\mathbf{HV} = \mathbf{D}$. Furthermore, $\langle\cdot\rangle_t$ indicates the mean over time of the expression included in the angle brackets.

In the next section we introduce the model system that we use for illustration throughout this letter. Section 3 describes two ways of analyzing a quadratic form by visualizing the coefficients of its quadratic and linear term directly and by considering the eigenvectors of its quadratic term. We then present in section 4 an algorithm to compute the optimal excitatory and inhibitory stimuli—the stimuli that maximize and minimize a quadratic form, respectively, given a fixed energy constraint. In section 5 we consider the invariances of the optimal stimuli, which are the transformations to which the function is most insensitive, and in the following section we introduce a test to determine which of these are statistically significant. In section 7 we discuss two ways to determine the relative contribution of the different terms of a quadratic form to its output. Furthermore, in section 8 we consider the techniques described above in the special case of a quadratic form without the linear term. In the end, we present in section 9 a two-layer neural network architecture equivalent to a given quadratic form. The letter concludes with a discussion of the relation of our approach to other studies in section 10.

2 Model System

To illustrate the analysis techniques presented here, we use the quadratic forms presented in Berkes and Wiskott (2002) in the context of a theoretical model of self-organization of complex-cell receptive fields in the primary visual cortex (see also Berkes & Wiskott, 2005b). In this section, we summarize the settings and main results of this example system.

We generated image sequences from a set of natural images by moving an input window over an image by translation, rotation, and zoom and subsequently rescaling the collected stimuli to a standard size of 16×16 pixels. For efficiency reasons, the dimensionality of the input vectors $\mathbf{x}$ was reduced from 256 to 50 input dimensions and whitened using principal component analysis (PCA). We then determined quadratic forms (also called *functions* or *units* in the following) by applying SFA to the input data. SFA is an implementation of the temporal slowness principle (see Wiskott & Sejnowski, 2002, and references there). Given a finite-dimensional function space, SFA extracts the functions that, applied to the input data, return output signals that vary as slowly as possible in time (as measured by the variance of the first derivative) under the constraint that the output signals have zero mean and unit variance and are decorrelated. The functions are sorted by decreasing slowness. For analysis, the quadratic forms are projected back from the 50 first principal components to the input space. Note that the rank of the quadratic term after the transformation is the same as before, and it thus has only 50 eigenvectors.

The units receive visual stimuli as an input and can be interpreted as nonlinear receptive fields. They were analyzed with the algorithms presented here and with sine-grating experiments similar to the ones performed in physiology and were found to reproduce many properties of complex cells in V1—not only the primary ones, that is, response to edges and phase-shift invariance (see sections 4 and 5), but also a range of secondary ones such as direction selectivity, nonorthogonal inhibition, end inhibition, and side inhibition.

This model system is complex enough to require an extensive analysis and is representative of the application domain considered here, which includes second-order approximations and theoretical models of physiological receptive fields.

3 Visualization of Coefficients and Eigenvectors

One way to analyze a quadratic form is to look at its coefficients. The coefficients $f_1, \ldots, f_N$ of the linear term can be visualized and interpreted directly. They give the shape of the input stimulus that maximizes the linear part given a fixed norm.

The quadratic term can be interpreted as a sum over the inner product of the jth row $\mathbf{h}_j$ of $\mathbf{H}$ with the vector of the products $x_j x_i$ between the jth

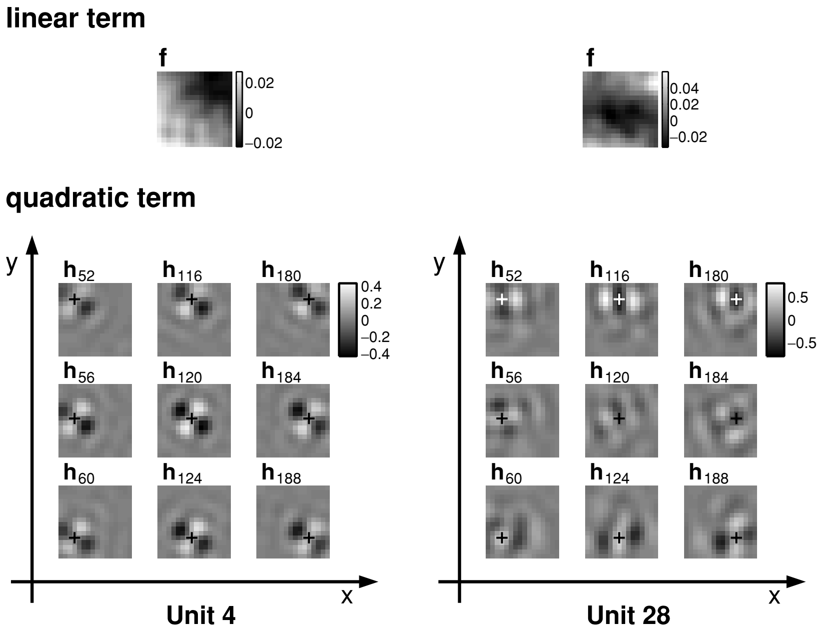

Figure 1: Some of the quadratic form coefficients of two functions learned in the model system. The top plots show the coefficients of the linear term **f**, reshaped to match the two-dimensional shape of the input. The bottom plots show the coefficients of nine of the rows $\mathbf{h}_j$ of the quadratic term. The crosses indicate the spatial position of the corresponding reference index j.

variable x_j and all other variables:

$$\mathbf{x}^T\mathbf{H}\mathbf{x} = \sum_{j=1}^{N} x_j(\mathbf{h}_j^T\mathbf{x}) = \sum_{j=1}^{N} \mathbf{h}_j^T \begin{pmatrix} x_j x_1 \\ x_j x_2 \\ \vdots \\ x_j x_N \end{pmatrix}. \tag{3.1}$$

In other words, the response of the quadratic term is formed by the sum of N linear filters $\mathbf{h}_j$ which respond to all combinations of the jth variable with the other ones.

If the input data have a two-dimensional spatial arrangement, as in our model system, the interpretation of the rows can be made easier by visualizing them as a series of images (by reshaping the vector $\mathbf{h}_j$ to match the structure of the input) and arranging them according to the spatial position of the corresponding variable x_j. In Figure 1 we show some of the

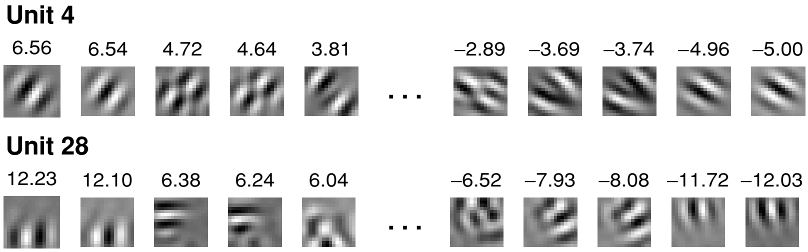

Figure 2: Eigenvectors of the quadratic term of two functions learned in the model system sorted by decreasing eigenvalues as indicated above each eigenvector.

coefficients of two units learned in the model system. In both, the linear term looks unstructured. The absolute values of its coefficients are small in comparison to those of the quadratic term so that its contribution to the output of the functions is very limited (cf. section 7). The row vectors $\mathbf{h}_j$ of unit 4 have a localized distribution of their coefficients; they respond only to combinations of the corresponding variable x_j and its neighbors. The filters $\mathbf{h}_j$ are shaped like a four-leaf clover and centered on the variable itself. Pairs of opposed leaves have positive and negative values, respectively. This suggests that the unit responds to stimuli oriented in the direction of the two positive leaves and is inhibited by stimuli with an orthogonal orientation, which is confirmed by successive analysis (cf. later in this section and section 4). In unit 28 the appearance of $\mathbf{h}_j$ depends on the spatial position of $\mathbf{x}_j$. In the bottom half of the receptive field, the interaction of the variables with their close neighbors along the vertical orientation is weighted positively, with a negative flank on the sides. In the top half, the rows have similar coefficients but with reversed polarity. As a consequence, the unit responds strongly to vertical edges in the bottom half, while vertical edges in the top half result in strong inhibition. Edges extending over the whole receptive field elicit only a weak total response. Such a unit is said to be end inhibited.

Another possibility for visualizing the quadratic term is to display its eigenvectors. The output of the quadratic form to one of the (normalized) eigenvectors equals half of the corresponding eigenvalue, since $\frac{1}{2}\mathbf{v}_i^T\mathbf{H}\mathbf{v}_i = \frac{1}{2}\mathbf{v}_i^T(\mu_i\mathbf{v}_i) = \frac{1}{2}\mu_i$. The first eigenvector can be interpreted as the stimulus that among all input vectors with norm 1 maximizes the output of the quadratic term. The jth eigenvector maximizes the quadratic term in the subspace that excludes the previous $j-1$ ones. In Figure 2 we show the eigenvectors of the two functions previously analyzed in Figure 1. In unit 4, the first eigenvector looks like a Gabor wavelet (i.e., a sine grating multiplied by a gaussian). The second eigenvector has the same form except for a 90 degree phase shift of the sine grating. Since the two eigenvalues have almost the same magnitude, the response of the quadratic term is similar for the two eigenvectors and also for linear combinations with constant norm 1. For this reason, the quadratic term of this unit has the main characteristics of complex cells in V1: a strong response to an oriented grating with an invariance to the phase of the grating. The last two eigenvectors, which correspond to the stimuli that minimize the quadratic term, are Gabor wavelets with orientation orthogonal to the first two. This means that the output of the quadratic term is inhibited by stimuli at an orientation orthogonal to the preferred one. A similar interpretation can be given in the case of unit 28, although in this case, the first and the last two eigenvalues have the same orientation but occupy two different halves of the receptive field. This confirms that unit 28 is end inhibited. A direct interpretation of the remaining eigenvectors in the two functions is difficult (see also section 8), although the magnitude of the eigenvalues shows that some of them elicit a strong response. Moreover, the interaction of the linear and quadratic terms to form the overall output of the quadratic form is not considered but cannot generally be neglected. The methods presented in the following sections often give a more direct and intuitive description of quadratic forms.

4 Optimal Stimuli

Another characterization of a nonlinear function can be borrowed from neurophysiological experiments, where it is common practice to characterize a neuron by the stimulus to which the neuron responds best (for an overview, see Dayan & Abbott, 2001). Analogously, we can compute the optimal excitatory stimulus of g, the input vector $\mathbf{x}^+$ that maximizes g given a fixed norm $\|\mathbf{x}^+\| = r$.[1] Note that $\mathbf{x}^+$ depends qualitatively on the value of r: if r is very small, the linear term of the equation dominates, so that $\mathbf{x}^+ \approx \mathbf{f}$, while if r is very large, the quadratic part dominates, so that $\mathbf{x}^+$ equals the first eigenvector of $\mathbf{H}$ (see also section 8). We usually choose r to be the mean norm of all input vectors, since we want $\mathbf{x}^+$ to be representative of the typical input. In the same way, we can also compute the optimal inhibitory stimulus $\mathbf{x}^-$, which minimizes the response of the function.

[1] The fixed norm constraint corresponds to a fixed energy constraint (Stork & Levinson, 1982) used in experiments involving the reconstruction of the Wiener kernel of a neuron (Dayan & Abbott, 2001). During physiological experiments in the visual system, one sometimes uses stimuli with fixed contrast instead. The optimal stimuli under these two constraints may be different. For example, with fixed contrast, one can extend a sine grating indefinitely in space without changing its intensity, while with fixed norm, its maximum intensity is going to dim as the extent of the grating increases. The fixed contrast constraint is more difficult to enforce analytically (e.g., because the surface of constant contrast is not bounded).

The problem of finding the optimal excitatory stimulus under the fixed energy constraint can be mathematically formulated as follows:

$$\begin{aligned} &\text{maximize} && g(\mathbf{x}) = \tfrac{1}{2}\mathbf{x}^T\mathbf{H}\mathbf{x} + \mathbf{f}^T\mathbf{x} + c \\ &\text{under the constraint} && \mathbf{x}^T\mathbf{x} = r^2 \,. \end{aligned} \tag{4.1}$$

This problem is known as the *trust region subproblem* and has been extensively studied in the context of numerical optimization, where a nonlinear function is minimized by successively approximating it by an inhomogeneous quadratic form, which is in turn minimized in a small neighborhood. Numerous studies have analyzed its properties in particular in the numerically difficult case where $\mathbf{H}$ is near to singular (see Fortin, 2000, and references there). We make use of some basic results and extend them where needed.

If the linear term is equal to zero (i.e., $\mathbf{f} = \mathbf{0}$), the problem can be easily solved (it is simply the first eigenvector scaled to norm r; see section 8). In the following, we consider the more general case where $\mathbf{f} \neq \mathbf{0}$. We can use a Lagrange formulation to find the necessary conditions for the extremum:

$$\mathbf{x}^T\mathbf{x} = r^2 \tag{4.2}$$

$$\text{and} \quad \nabla[g(\mathbf{x}) - \tfrac{1}{2}\lambda\mathbf{x}^T\mathbf{x}] = \mathbf{0} \tag{4.3}$$

$$\Leftrightarrow \quad \mathbf{H}\mathbf{x} + \mathbf{f} - \lambda\mathbf{x} = \mathbf{0} \tag{4.4}$$

$$\Leftrightarrow \quad \mathbf{x} = (\lambda\mathbf{I} - \mathbf{H})^{-1}\mathbf{f} \,, \tag{4.5}$$

where we inserted the factor $\frac{1}{2}$ for mathematical convenience. According to theorem 3.1 in Fortin (2000), if an $\mathbf{x}$ that satisfies equation 4.5 is a solution to equation 4.1, then $(\lambda\mathbf{I} - \mathbf{H})$ is positive semidefinite (i.e., all eigenvalues are greater than or equal to 0). This imposes a strong lower bound on the range of possible values for λ. Note that the matrix $(\lambda\mathbf{I} - \mathbf{H})$ has the same eigenvectors $\mathbf{v}_i$ as $\mathbf{H}$ with eigenvalues $(\lambda - \mu_i)$. For $(\lambda\mathbf{I} - \mathbf{H})$ to be positive semidefinite, all eigenvalues must be nonnegative, and thus λ must be greater than the largest eigenvalue μ_1,

$$\mu_1 \leq \lambda \,. \tag{4.6}$$

An upper bound for lambda can be found by considering an upper bound for the norm of $\mathbf{x}$. First, we note that matrix $(\lambda\mathbf{I} - \mathbf{H})^{-1}$ is symmetric and has the same eigenvectors as $\mathbf{H}$ with eigenvalues $1/(\lambda - \mu_i)$. We also know that $\|\mathbf{A}\mathbf{v}\| \leq \|\mathbf{A}\|\|\mathbf{v}\|$ for every matrix $\mathbf{A}$ and vector $\mathbf{v}$. $\|\mathbf{A}\|$ is here the spectral norm of $\mathbf{A}$, which for symmetric matrices is simply the largest absolute eigenvalue. With this we find an upper bound for λ:

$$\begin{aligned} r &= \|\mathbf{x}\| && (4.7)\\ &= \|(\lambda\mathbf{I}-\mathbf{H})^{-1}\mathbf{f}\| && (4.8)\\ &\leq \|(\lambda\mathbf{I}-\mathbf{H})^{-1}\|\ \|\mathbf{f}\| && (4.9)\\ &= \max_i \left\{\left|\frac{1}{\lambda-\mu_i}\right|\right\}\ \|\mathbf{f}\| && (4.10)\\ &\underset{(4.6)}{=} \frac{1}{\lambda-\mu_1}\ \|\mathbf{f}\| && (4.11)\\ \Leftrightarrow \quad \lambda &\leq \frac{\|\mathbf{f}\|}{r}+\mu_1\,. && (4.12)\end{aligned}$$

The optimization problem, equation 4.1, is thus reduced to a search over λ on the interval $[\mu_1, (\frac{\|\mathbf{f}\|}{r}+\mu_1)]$ until $\mathbf{x}$ defined by equation 4.5 fulfills the constraint $\|\mathbf{x}\| = r$ (see equation 4.2). Vector $\mathbf{x}$ and norm $\|\mathbf{x}\|$ can be efficiently computed for each λ using the eigenvalue decomposition of $\mathbf{f}$:

$$\begin{aligned} \mathbf{x} &\underset{(4.5)}{=} (\lambda\mathbf{I}-\mathbf{H})^{-1}\mathbf{f} && (4.13)\\ &= (\lambda\mathbf{I}-\mathbf{H})^{-1}\sum_i \mathbf{v}_i\,(\mathbf{v}_i^T\mathbf{f}) && (4.14)\\ &= \sum_i (\lambda\mathbf{I}-\mathbf{H})^{-1}\ \mathbf{v}_i\,(\mathbf{v}_i^T\mathbf{f}) && (4.15)\\ &= \sum_i \frac{1}{\lambda-\mu_i}\mathbf{v}_i\ (\mathbf{v}_i^T\mathbf{f}) && (4.16)\end{aligned}$$

and

$$\|\mathbf{x}\|^2 = \sum_i \left(\frac{1}{\lambda-\mu_i}\right)^2 (\mathbf{v}_i^T\mathbf{f})^2\,, \qquad (4.17)$$

where the terms $\mathbf{v}_i^T\mathbf{f}$ and $(\mathbf{v}_i^T\mathbf{f})^2$ are constant for each quadratic form and can be computed in advance. The last equation also shows that the norm of $\mathbf{x}$ is monotonically decreasing in the considered interval, so that there is exactly one solution and the search can be efficiently performed by a bisection method. $\mathbf{x}^-$ can be found in the same way by maximizing the negative of g. The pseudocode of an algorithm that implements all the considerations above can be found in Berkes and Wiskott (2005a). A Matlab version can be downloaded online from the authors' home pages (http://itb.biologie.hu-berlin.de/{˜berkes,˜wiskott}).

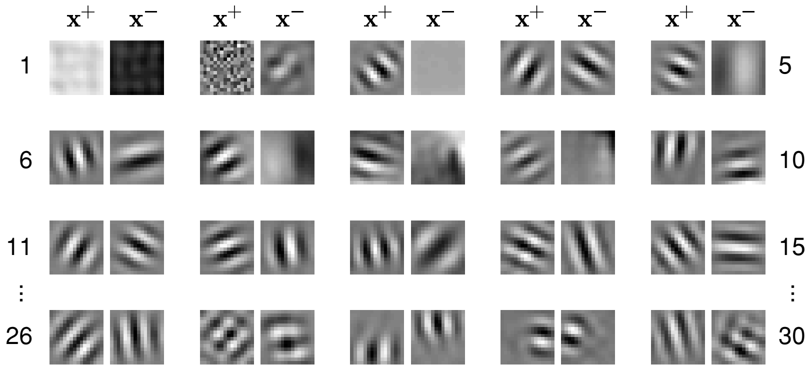

Figure 3: Optimal stimuli of some of the units in the model system. $\mathbf{x}^+$ looks like a Gabor wavelet in almost all cases, in agreement with physiological data. $\mathbf{x}^-$ is usually structured and is also similar to a Gabor wavelet, which suggests that inhibition plays an important role.

If the matrix $\mathbf{H}$ is negative definite (i.e., all its eigenvalues are negative) there is a global maximum that may not lie on the sphere, which might be used in substitution for $\mathbf{x}^+$ if it lies in a region of the input space that has a high probability of being reached (the criterion is quite arbitrary, but the region could be chosen to include, for example, 75% of the input data with highest density). The gradient of the function disappears at the global extremum such that it can be found by solving a simple linear equation system:

$$\nabla g(\mathbf{x}) = \mathbf{H}\mathbf{x} + \mathbf{f} = \mathbf{0} \tag{4.18}$$

$$\Leftrightarrow \quad \mathbf{x} = -\mathbf{H}^{-1}\mathbf{f}. \tag{4.19}$$

In the same way, a positive definite matrix $\mathbf{H}$ has a negative global minimum, which might be used in substitution for $\mathbf{x}^-$.

In Figure 3 we show the optimal stimuli of some of the units in the model system. In almost all cases, $\mathbf{x}^+$ looks like a Gabor wavelet, in agreement with physiological data for neurons of the primary visual cortex (Pollen & Ronner, 1981; Adelson & Bergen, 1985; Jones & Palmer, 1987). The functions respond best to oriented stimuli having the same frequency as $\mathbf{x}^+$. $\mathbf{x}^-$ is usually structured as well and looks like a Gabor wavelet too, which suggests that inhibition plays an important role. $\mathbf{x}^+$ can be used to compute the position and size of the receptive fields as well as the preferred orientation and frequency of the units for successive experiments.

Note that although $\mathbf{x}^+$ is the stimulus that elicits the strongest response in the function, it does not necessarily mean that it is representative of the class of stimuli that give the most important contribution to its output. This depends on the distribution of the input vectors. If $\mathbf{x}^+$ lies in a low-density region of the input space, it is possible that other kinds of stimuli drive the function more often. In that case, they might be considered more relevant than $\mathbf{x}^+$ to characterize the function. Symptomatic for this effect would be if the output of a function when applied to its optimal stimulus would lie far outside the range of normal activity. This means that $\mathbf{x}^+$ can be an atypical, artificial input that pushes the function in an uncommon state. A similar effect has also been reported in a physiological article comparing the response of neurons to natural stimuli and to artificial stimuli such as sine gratings (Baddeley et al., 1997). The characterization of a neuron or a nonlinear function as a feature detector via the optimal stimulus is thus at least incomplete (see also MacKay, 1985). However, the optimal stimuli remain extremely informative in practice.

5 Invariances

Since the considered functions are nonlinear, the optimal stimuli do not provide a complete description of their properties. We can gain some additional insights by studying a neighborhood of $\mathbf{x}^+$ and $\mathbf{x}^-$. An interesting question is to which transformations of $\mathbf{x}^+$ or $\mathbf{x}^-$ the function is invariant. This is similar to the common interpretation of neurons as detectors of a specific feature of the input that are invariant to a local transformation of that feature. For example, complex cells in the primary visual cortex are thought to respond to oriented bars and to be invariant to a local translation. In this section, we consider the function $\tilde{g}$ defined as g restricted to the sphere $\mathcal{S}$ of radius r, since as in section 4, we want to compare input vectors having fixed energy. Notice that although $\tilde{g}$ and g take the same values on $\mathcal{S}$ (i.e., $\tilde{g}(\mathbf{x}) = g(\mathbf{x})$ for each $\mathbf{x} \in \mathcal{S}$), they are two distinct mathematical objects. For example, the gradient of $\tilde{g}$ in $\mathbf{x}^+$ is zero because $\mathbf{x}^+$ is by definition a maximum of $\tilde{g}$. On the other hand, the gradient of g in the same point is $\mathbf{H}\mathbf{x}^+ + \mathbf{f}$, which is in general different from zero.

Strictly speaking, there is no invariance in $\mathbf{x}^+$, since it is a maximum, and the output of $\tilde{g}$ decreases in all directions (except in the special case where the linear term is zero and the first two or more eigenvalues are equal). On the other hand, in a general, noncritical point $\mathbf{x}^*$ (i.e., a point where the gradient does not disappear), the rate of change in any direction $\mathbf{w}$ is given by its inner product with the gradient, $\nabla\tilde{g}(\mathbf{x}^*) \cdot \mathbf{w}$. For all vectors orthogonal to the gradient (which span an $N - 2$ dimensional space), the rate of change is thus zero. Note that this is not merely a consequence of the fact that the gradient is a first-order approximation of $\tilde{g}$. By the implicit function theorem (see e.g., Walter, 1995, theorem 4.5), in each open neighborhood U of a noncritical point $\mathbf{x}^*$, there is an $N - 2$–dimensional level surface $\{\mathbf{x} \in U \subset \mathcal{S} \mid \tilde{g}(\mathbf{x}) = \tilde{g}(\mathbf{x}^*)\}$, since the domain of $\tilde{g}$ (the sphere $\mathcal{S}$) is an $N - 1$–dimensional surface and its range ($\mathbb{R}$) is one-dimensional. Each noncritical point thus belongs to an $N - 2$ dimensional surface where

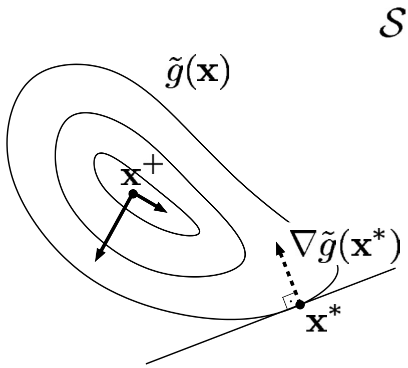

Figure 4: Definition of invariance. This figure shows a contour plot of $\tilde{g}(\mathbf{x})$ on the surface of the sphere $\mathcal{S}$ in a neighborhood of $\mathbf{x}^+$. Each general point $\mathbf{x}^*$ on $\mathcal{S}$ lies on an $N-2$–dimensional level surface (as indicated by the closed lines) where the output of the function $\tilde{g}$ does not change. The only interesting direction in $\mathbf{x}^*$ is the one of maximal change, as indicated by the gradient $\nabla\tilde{g}(\mathbf{x}^*)$. On the space orthogonal to it, the rate of change is zero. In $\mathbf{x}^+$ the function has a maximum, and its output decreases in all directions. There is thus no strict invariance. Considering the second derivative, however, we can identify the directions of minimal change. The arrows in $\mathbf{x}^+$ indicate the direction of the invariances (see equation 5.9) with a length proportional to the corresponding second derivative.

the value of the $\tilde{g}$ stays constant. This is a somewhat surprising result: for an optimal stimulus, there does not exist any invariance (except in some degenerate cases); for a general suboptimal stimulus, there exist many invariances.

This shows that although it might be useful to observe, for example, that a given function f that maps images to real values is invariant to stimulus rotation, one should keep in mind that in a generic point, there is a large number of other transformations to which the function is equally invariant but would lack an easy interpretation. The strict concept of invariance is thus not useful for our analysis, since in the extrema we have no invariances at all, while in a general point, they are the typical case and the only interesting direction is the one of maximal change, as indicated by the gradient. In the extremum $\mathbf{x}^+$, however, since the output changes in all directions, we can relax the definition of invariance and look for the transformation to which the function changes as little as possible, as indicated by the direction with the smallest absolute value of the second derivative (see Figure 4). (In a noncritical point, this weak definition of invariance still does not help. If the quadratic form that represents the second derivative has positive as well as negative eigenvalues, there is still an $N-3$–dimensional surface where the second derivative is zero.)

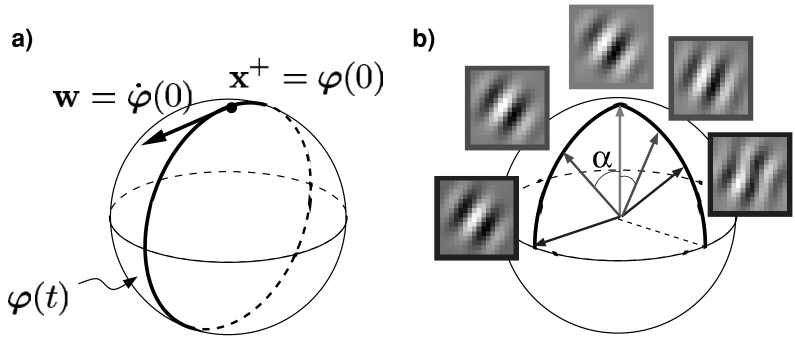

Figure 5: Invariances. (a) To compute the second derivative of the quadratic form on the surface of the sphere, one can study the function along special paths on the sphere, known as geodetics. Geodetics of a sphere are great circles. (b) This plot illustrates how the invariances are visualized. Starting from the optimal stimulus (top), we move on the sphere in the direction of an invariance until the response of the function drops below 80% of the maximal output or α reaches 90 degrees. In the figure, two invariances of unit 4 are visualized. The one on the left represents a phase shift invariance and preserves more than 80% of the maximal output until 90 degrees (the output at 90 degrees is 99.6% of the maximum). The one on the right represents an invariance to orientation change with an output that drops below 80% at 55 degrees.

To study the invariances of the function g in a neighborhood of its optimal stimulus respecting the fixed energy constraint, we have defined the function $\tilde{g}$ as the function g restricted to $\mathcal{S}$. This is particularly relevant here since we want to analyze the derivatives of the function, that is, its change under small movements. Any straight movement in space is going to leave the surface of the sphere. We must therefore be able to define movements on the sphere itself. This can be done by considering a path $\boldsymbol{\varphi}(t)$ on the surface of $\mathcal{S}$ such that $\boldsymbol{\varphi}(0) = \mathbf{x}^+$ and then studying the change of g along $\boldsymbol{\varphi}$. By doing this, however, we add the rate of change of the path (i.e., its acceleration) to that of the function. Of all possible paths, we must take the ones that have as little acceleration as possible—those that have just the acceleration that is needed to stay on the surface. Such a path is called a *geodetic*. The geodetics of a sphere are great circles, and our paths are thus defined as

$$\boldsymbol{\varphi}(t) = \cos(t/r) \cdot \mathbf{x}^+ + \sin(t/r) \cdot r\mathbf{w} \tag{5.1}$$

for each direction $\mathbf{w}$ in the tangential space of $\mathcal{S}$ in $\mathbf{x}^+$ (i.e., for each $\mathbf{w}$ orthogonal to $\mathbf{x}^+$), as shown in Figure 5a. The $1/r$ factor in the cosine and sine arguments normalizes the function such that $\frac{\mathrm{d}}{\mathrm{d}t}\boldsymbol{\varphi}(0) = \mathbf{w}$ with $\|\mathbf{w}\| = 1$.

For the first derivative of $\tilde{g}$ along $\boldsymbol{\varphi}$, we obtain by straightforward calculations (with $(\tilde{g} \circ \boldsymbol{\varphi})(t) := \tilde{g}(\boldsymbol{\varphi}(t))$)

$$\frac{\mathrm{d}}{\mathrm{d}t}(\tilde{g} \circ \boldsymbol{\varphi})(t) = \frac{\mathrm{d}}{\mathrm{d}t}\left[\frac{1}{2}\boldsymbol{\varphi}(t)^T \mathbf{H}\boldsymbol{\varphi}(t) + \mathbf{f}^T \boldsymbol{\varphi}(t) + c\right] = \dots \tag{5.2}$$

$$\begin{aligned} &= -\frac{1}{r}\sin(t/r)\cos(t/r)\; \mathbf{x}^{+T}\mathbf{H}\mathbf{x}^{+} + \cos(2t/r)\; \mathbf{x}^{+T}\mathbf{H}\mathbf{w} \\ &\quad + \sin(t/r)\cos(t/r)\; r\; \mathbf{w}^T\mathbf{H}\mathbf{w} \\ &\quad - \frac{1}{r}\sin(t/r)\; \mathbf{f}^T\mathbf{x}^{+} + \cos(t/r)\; \mathbf{f}^T\mathbf{w}\,, \end{aligned} \tag{5.3}$$

and for the second derivative,

$$\begin{aligned} \frac{\mathrm{d}^2}{\mathrm{d}t^2}(\tilde{g} \circ \boldsymbol{\varphi})(t) &= -\frac{1}{r^2}\cos(2t/r)\; \mathbf{x}^{+T}\mathbf{H}\mathbf{x}^{+} - \frac{2}{r}\sin(2t/r)\; \mathbf{x}^{+T}\mathbf{H}\mathbf{w} \\ &\quad + \cos(2t/r)\; \mathbf{w}^T\mathbf{H}\mathbf{w} - \frac{1}{r^2}\cos(t/r)\; \mathbf{f}^T\mathbf{x}^{+} - \frac{1}{r}\sin(t/r)\; \mathbf{f}^T\mathbf{w}\,. \end{aligned} \tag{5.4}$$

In $t = 0$ we have

$$\frac{\mathrm{d}^2}{\mathrm{d}t^2}(\tilde{g} \circ \boldsymbol{\varphi})(0) = \mathbf{w}^T\mathbf{H}\mathbf{w} - \frac{1}{r^2}(\mathbf{x}^{+T}\mathbf{H}\mathbf{x}^{+} + \mathbf{f}^T\mathbf{x}^{+})\,, \tag{5.5}$$

that is, the second derivative of $\tilde{g}$ in $\mathbf{x}^{+}$ in the direction of $\mathbf{w}$ is composed of two terms: $\mathbf{w}^T\mathbf{H}\mathbf{w}$ corresponds to the second derivative of g in the direction of $\mathbf{w}$, while the constant term $-1/r^2 \cdot (\mathbf{x}^{+T}\mathbf{H}\mathbf{x}^{+} + \mathbf{f}^T\mathbf{x}^{+})$ depends on the curvature of the sphere $1/r^2$ and on the gradient of g in $\mathbf{x}^{+}$ orthogonal to the surface of the sphere,

$$\nabla g(\mathbf{x}^{+}) \cdot \mathbf{x}^{+} = (\mathbf{H}\mathbf{x}^{+} + \mathbf{f})^T\mathbf{x}^{+} \tag{5.6}$$

$$= \mathbf{x}^{+T}\mathbf{H}\mathbf{x}^{+} + \mathbf{f}^T\mathbf{x}^{+}. \tag{5.7}$$

To find the direction in which $\tilde{g}$ decreases as little as possible, we only need to minimize the absolute value of the second derivative (see equation 5.5). This is equivalent to maximizing the first term $\mathbf{w}^T\mathbf{H}\mathbf{w}$ in equation 5.5 since the second derivative in $\mathbf{x}^{+}$ is always negative (because $\mathbf{x}^{+}$ is a maximum of $\tilde{g}$) and the second term is constant. $\mathbf{w}$ is orthogonal to $\mathbf{x}^{+}$, and thus the maximization must be performed in the space tangential to the sphere in $\mathbf{x}^+$. This can be done by computing a basis $\mathbf{b}_2, \ldots, \mathbf{b}_N$ of the tangential space (e.g., using the Gram-Schmidt orthogonalization on $\mathbf{x}^+, \mathbf{e}_1, \ldots, \mathbf{e}_{N-1}$ where $\mathbf{e}_i$ is the canonical basis of $\mathbb{R}^N$) and replacing the matrix $\mathbf{H}$ by

$$\tilde{\mathbf{H}} = \mathbf{B}^T \mathbf{H} \mathbf{B}\,, \tag{5.8}$$

where $\mathbf{B} = (\mathbf{b}_2, \cdots, \mathbf{b}_N)$. The direction of the smallest second derivative corresponds to the eigenvector $\tilde{\mathbf{v}}_1$ of $\tilde{\mathbf{H}}$ with the largest positive eigenvalue. The eigenvector must then be projected back from the tangential space into the original space by a multiplication with $\mathbf{B}$:

$$\mathbf{w}_1 = \mathbf{B}\tilde{\mathbf{v}}_1\,. \tag{5.9}$$

The remaining eigenvectors corresponding to eigenvalues of decreasing value are also interesting, as they point in orthogonal directions where the function changes with a gradually increasing rate of change.

To visualize the invariances, we move $\mathbf{x}^+$ (or $\mathbf{x}^-$) along a path on the sphere in the direction of a vector $\mathbf{w}_i$ according to

$$\mathbf{x}(\alpha) = \cos(\alpha) \cdot \mathbf{x}^+ + \sin(\alpha) \cdot r\mathbf{w}_i \tag{5.10}$$

for $\alpha \in [-90^\circ, 90^\circ]$, as illustrated in Figure 5b. At each point, we measure the response of the function to the new input vector and stop when it drops below 80% of the maximal response. In this way, we generate for each invariance a movie like those shown in Figure 6 for some of the optimal stimuli (the corresponding animations are available at the authors' home pages). Each frame of such a movie contains a nearly optimal stimulus. Using this analysis, we can systematically scan a neighborhood of the optimal stimuli, starting from the transformations to which the function is most insensitive up to those that lead to a great change in response. Note that our definition of invariance applies only locally to a small neighborhood of $\mathbf{x}^+$. The path followed in equation 5.10 goes beyond such a neighborhood and is appropriate only for visualization. The pseudocode of an algorithm that computes and visualizes the invariances of the optimal stimuli can be found in Berkes and Wiskott (2005a). A Matlab version can be downloaded from the authors' home pages.

6 Significant Invariances

The procedure described above finds for each optimal stimulus a set of $N-1$ invariances ordered by the degree of invariance (i.e., by increasing magnitude of the second derivative). We would like to know which of these are statistically significant. An invariance can be defined as significant

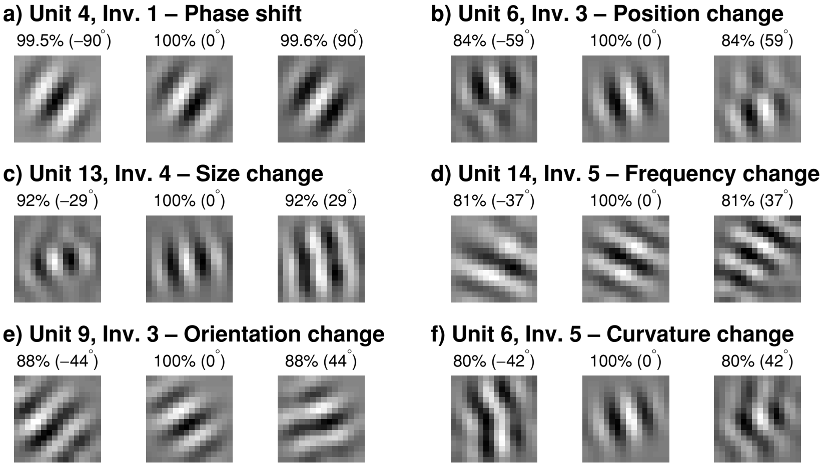

Figure 6: Selected invariances for some of the optimal excitatory stimuli shown in Figure 3. The central patch of each plot represents the optimal stimulus of a unit, while the ones on the sides are produced by moving it in one (left patch) or the other (right patch) direction of the eigenvector corresponding to the invariance. In this image, we stopped before the output dropped below 80% of the maximum to make the interpretation of the invariances easier. The relative output of the function in percent and the angle of displacement α (see equation 5.10) are given above the patches. The animations corresponding to these invariances are available at the authors' home pages.

if the function changes exceptionally little (less than chance level) in that direction, which can be measured by the value of the second derivative: the smaller its absolute value, the slower the function will change.

To test for their significance, we compare the second derivatives of the invariances of the quadratic form we are considering with those of random inhomogeneous quadratic forms that are equally adapted to the statistics of the input data. We therefore constrain the random quadratic forms to produce an output that has the same variance and mean as the output of the analyzed ones when applied to the input stimuli. Without loss of generality, we assume here zero mean and unit variance. These constraints are compatible with the ones that are usually imposed on the functions learned by many theoretical models. Because of this normalization, the distribution of the random quadratic forms depends on the distribution of the input data.

To understand how to efficiently build random quadratic forms under these constraints, it is useful to think in terms of a dual representation of the problem. A quadratic form over the input space is equivalent to a linear function over the space of the input expanded to all monomials of degree one and two using the function $\boldsymbol{\Phi}((x_1, \ldots, x_n)^T) := (x_1x_1, x_1x_2, x_1x_3, \ldots, x_nx_n, x_1, \ldots, x_n)^T$, that is,

$$\frac{1}{2}\mathbf{x}^T \underbrace{\begin{pmatrix} h_{11} & h_{12} & \cdots & h_{1n} \\ h_{12} & h_{22} & & \\ \vdots & & \ddots & \vdots \\ h_{1n} & & \cdots & h_{nn} \end{pmatrix}}_{\mathbf{H}} \mathbf{x} + \underbrace{\begin{pmatrix} f_1 \\ f_2 \\ \vdots \\ f_n \end{pmatrix}}_{\mathbf{f}}^T \mathbf{x} + c = \underbrace{\begin{pmatrix} \frac{1}{2}h_{11} \\ h_{12} \\ h_{13} \\ \vdots \\ \frac{1}{2}h_{nn} \\ f_1 \\ \vdots \\ f_n \end{pmatrix}}_{\mathbf{q}}^T \underbrace{\begin{pmatrix} x_1x_1 \\ x_1x_2 \\ x_1x_3 \\ \vdots \\ x_nx_n \\ x_1 \\ \vdots \\ x_n \end{pmatrix}}_{\Phi(\mathbf{x})} + c\,. \tag{6.1}$$

We can whiten the expanded input data $\boldsymbol{\Phi}(\mathbf{x})$ by subtracting its mean $\langle\boldsymbol{\Phi}(\mathbf{x})\rangle_t$ and transforming it with a whitening matrix $\mathbf{S}$. In this new coordinate system, each vector with norm 1 applied to the input data using the scalar product fulfills the unit variance and zero mean constraints by construction. We can thus choose a random vector $\mathbf{q}'$ of length 1 in the whitened, expanded space and derive the corresponding quadratic form in the original input space:

$$\mathbf{q}'^T\left(\mathbf{S}(\boldsymbol{\Phi}(\mathbf{x}) - \langle\boldsymbol{\Phi}(\mathbf{x})\rangle_t)\right) = \underbrace{(\mathbf{S}^T\mathbf{q}')}_{=:\mathbf{q}}{}^T\left(\boldsymbol{\Phi}(\mathbf{x}) - \langle\boldsymbol{\Phi}(\mathbf{x})\rangle_t\right) \tag{6.2}$$

$$= \mathbf{q}^T(\boldsymbol{\Phi}(\mathbf{x}) - \langle\boldsymbol{\Phi}(\mathbf{x})\rangle_t) \tag{6.3}$$

$$\underset{(6.1)}{=} \frac{1}{2}\mathbf{x}^T\mathbf{H}\mathbf{x} + \mathbf{f}^T\mathbf{x} - \underbrace{\mathbf{q}^T\langle\boldsymbol{\Phi}(\mathbf{x})\rangle_t}_{:=c} \tag{6.4}$$

$$= \frac{1}{2}\mathbf{x}^T\mathbf{H}\mathbf{x} + \mathbf{f}^T\mathbf{x} + c\,, \tag{6.5}$$

with appropriately defined $\mathbf{H}$ and $\mathbf{f}$ according to equation 6.1.

We can next compute the optimal stimuli and the second derivative of the invariances of the obtained random quadratic form. To make sure that we get independent measurements, we keep only one second derivative chosen at random for each random function. This operation, repeated over many quadratic forms, allows us to determine a distribution of the second derivatives of the invariances and a corresponding confidence interval.

Figure 7a shows the distribution of 50,000 independent second derivatives of the invariances of random quadratic forms and the distribution of the second derivatives of all invariances of the first 50 units learned in the model system. The dashed line indicates the 95% confidence interval

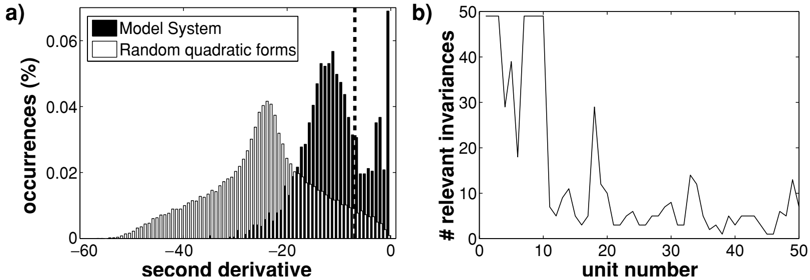

Figure 7: Significant invariances. (a) Distribution of 50,000 independently drawn second derivatives of the invariances of random quadratic forms and distribution of the second derivatives of all invariances of the first 50 units learned in the model system. The dashed line indicates the 95% confidence interval as derived from the random quadratic forms. The distribution in the model system is more skewed toward small second derivatives and has a clear peak near zero. Twenty-eight percent of all invariances were classified as significant. (b) Number of significant invariances for each of the first 50 units learned in the model system (the functions were sorted by decreasing slowness; see section 2). The number of significant invariances tends to decrease with decreasing slowness.

derived from the former distribution. The latter is more skewed toward small second derivatives and has a clear peak near zero. Twenty-eight percent of all invariances were classified as significant. Figure 7b shows the number of significant invariances for each individual quadratic form in the model system. Each function has 49 invariances since the rank of the quadratic term is 50 (see section 2). The plot shows that the number of significant invariances decreases with increasing ordinal number (the functions are ordered by slowness, the first ones being the slowest). Forty-six units out of 50 have three or more significant invariances. The first invariance, which corresponds to a phase shift invariance, was always classified as significant, which confirms that the units behave like complex cells. Note that since the eigenvalues of a quadratic form are not independent of each other, with the method presented here it is possible to make statements only about the significance of individual invariances, and not about the joint probability distribution of two or more invariances.

7 Relative Contribution of the Quadratic, Linear, and Constant Term

The inhomogeneous quadratic form has a quadratic, a linear, and a constant term. It is sometimes of interest to know what their relative contribution to the output is. The answer to this question depends on the considered input. For example, the quadratic term dominates for large input vectors,

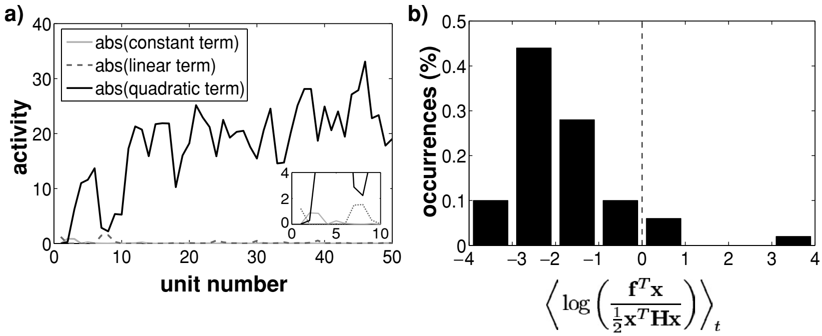

Figure 8: Relative contribution of the quadratic, linear, and constant term. (a) The absolute value of the output of the quadratic, linear, and constant term in $\mathbf{x}^+$ for each of the first 50 units in the model system. In all but the first 2 units, the quadratic term has a larger output. The subplot shows a magnified version of the contribution of the terms for the first 10 units. (b) Histogram of the mean of the logarithm of the ratio between the activity of the linear and the quadratic term in the model system when applied to 90,000 test input vectors. A negative value means that the quadratic term dominates, and a positive value means the linear term dominates. In all but 4 units (units 1, 7, 8, and 24), the quadratic term is greater on average.

while the linear or even the constant term dominates for input vectors with a small norm.

A first possibility is to look at the contribution of the individual terms at a particular point. A privileged point is, for example, the optimal excitatory stimulus, especially if the quadratic form can be interpreted as a feature detector (cf. section 4). Figure 8a shows for each function in the model system the absolute value of the output of all terms with $\mathbf{x}^+$ as an input. In all functions except the first two, the activity of the quadratic term is greater than that of the linear and of the constant term. The first function basically computes the mean pixel intensity, which explains the dominance of the linear term. The second function is dominated by a constant term from which a quadratic expression very similar to the squared mean pixel intensity is subtracted.

As an alternative we can consider the ratio between linear and quadratic term, averaged over all input stimuli:

$$\left\langle \log\left|\frac{\mathbf{f}^T\mathbf{x}}{\frac{1}{2}\mathbf{x}^T\mathbf{H}\mathbf{x}}\right|\right\rangle_t = \left\langle \log\left|\mathbf{f}^T\mathbf{x}\right| - \log\left|\frac{1}{2}\mathbf{x}^T\mathbf{H}\mathbf{x}\right|\right\rangle_t . \tag{7.1}$$

The logarithm ensures that a given ratio (e.g., linear/quadratic = 3) has the same weight as the inverse ratio (e.g., linear/quadratic = 1/3) in the mean.

A negative result means that the quadratic term dominates, while for a positive value, the linear term dominates. Figure 8b shows the histogram of this measure for the functions in the model syatem. In all but 4 units (units 1, 7, 8, and 24), the quadratic term is on average greater than the linear one.

8 Quadratic Forms Without the Linear Term

In the case of a quadratic form without the linear term,

$$g(\mathbf{x}) = \frac{1}{2}\mathbf{x}^T\mathbf{H}\mathbf{x} + c, \tag{8.1}$$

the mathematics of sections 4 and 5 becomes much simpler. The quadratic form is now centered at $\mathbf{x} = \mathbf{0}$, and the direction of maximal increase corresponds to the eigenvector $\mathbf{v}_1$ with the largest positive eigenvalue. The optimal excitatory stimulus $\mathbf{x}^+$ with norm r is thus

$$\mathbf{x}^+ = r\mathbf{v}_1\,. \tag{8.2}$$

Similarly, the eigenvector corresponding to the largest negative eigenvalue $\mathbf{v}_N$ points in the direction of $\mathbf{x}^-$.

The second derivative, equation 5.5, in $\mathbf{x}^+$ in this case becomes

$$\frac{\mathrm{d}^2}{\mathrm{d}t^2}(\tilde{g}\circ\boldsymbol{\varphi})(0) = \mathbf{w}^T\mathbf{H}\mathbf{w} - \frac{1}{r^2}\mathbf{x}^{+T}\mathbf{H}\mathbf{x}^+ \tag{8.3}$$

$$\underset{(8.2)}{=} \mathbf{w}^T\mathbf{H}\mathbf{w} - \mathbf{v}_1^T\mathbf{H}\mathbf{v}_1 \tag{8.4}$$

$$= \mathbf{w}^T\mathbf{H}\mathbf{w} - \mu_1. \tag{8.5}$$

The vector $\mathbf{w}$ is by construction orthogonal to $\mathbf{x}^+$ and therefore lies in the space spanned by the remaining eigenvectors $\mathbf{v}_2, \ldots, \mathbf{v}_N$. Since μ_1 is the maximum value that $\mathbf{w}^T\mathbf{H}\mathbf{w}$ can assume for vectors of length 1, it is clear that equation 8.5 is always negative (as it should since $\mathbf{x}^+$ is a maximum) and that its absolute value is successively minimized by the eigenvectors $\mathbf{v}_2, \ldots, \mathbf{v}_N$ in this order. The value of the second derivative on the sphere in the direction of $\mathbf{v}_i$ is given by

$$\frac{\mathrm{d}^2}{\mathrm{d}t^2}(\tilde{g}\circ\boldsymbol{\varphi})(0) = \mathbf{v}_i^T\mathbf{H}\mathbf{v}_i - \mu_1 \tag{8.6}$$

$$= \mu_i - \mu_1\,. \tag{8.7}$$

In the same way, the invariances of $\mathbf{x}^-$ are given by $\mathbf{v}_{N-1}, \ldots, \mathbf{v}_1$ with second derivative values $(\mu_i - \mu_N)$.

Since, as shown in Figure 8a, in the model system the linear term is mostly small in comparison with the quadratic one, the first and last eigenvectors of our units are expected to be very similar to their optimal stimuli. This can be verified by comparing Figures 2 and 3. Moreover, successive eigenvectors are almost equal to the directions of the most relevant invariances (compare, for example, unit 4 in Figure 2 and Figure 5b). This does not have to be the case in general. For example, the data in Lewis et al. (2002) show that cochlear neurons in the gerbil ear have a linear as well as a quadratic component. In such a situation, the algorithms must be applied in their general formulation.

9 Decomposition of a Quadratic Form in a Neural Network

As also noticed by Hashimoto (2003), for each quadratic form there exists an equivalent two-layer neural network, which can be derived by rewriting the quadratic form using its eigenvector decomposition:

$$g(\mathbf{x}) = \frac{1}{2}\mathbf{x}^T\mathbf{H}\mathbf{x} + \mathbf{f}^T\mathbf{x} + c \tag{9.1}$$

$$= \frac{1}{2}\mathbf{x}^T\mathbf{V}\mathbf{D}\mathbf{V}^T\mathbf{x} + \mathbf{f}^T\mathbf{x} + c \tag{9.2}$$

$$= \frac{1}{2}(\mathbf{V}^T\mathbf{x})^T\mathbf{D}(\mathbf{V}^T\mathbf{x}) + \mathbf{f}^T\mathbf{x} + c \tag{9.3}$$

$$= \sum_{i=1}^{N}\frac{\mu_i}{2}(\mathbf{v}_i^T\mathbf{x})^2 + \mathbf{f}^T\mathbf{x} + c\,. \tag{9.4}$$

The network has a first layer formed by a set of N linear subunits $s_k(\mathbf{x}) = \mathbf{v}_k^T\mathbf{x}$ followed by a quadratic nonlinearity weighted by the coefficients $\mu_k/2$. The output neuron sums the contribution of all subunits plus the output of a direct linear connection from the input layer (see Figure 9a). Since the eigenvalues can be negative, some of the subunits give an inhibitory contribution to the output. It is interesting to note that in an algorithm that learns quadratic forms, the number of inhibitory subunits in the equivalent neural network is not fixed but is a learned feature. As an alternative, one can scale the weights $\mathbf{v}_i$ by $\sqrt{|\mu_i|/2}$ and specify which subunits are excitatory and which are inhibitory according to the sign of μ_i, since

$$g(\mathbf{x}) \underset{(9.4)}{=} \sum_{i=1}^{N}\frac{\mu_i}{2}(\mathbf{v}_i^T\mathbf{x})^2 + \mathbf{f}^T\mathbf{x} + c \tag{9.5}$$

$$= \sum_{\substack{i=1\\ \mu_i>0}}^{N}\left(\left(\sqrt{\frac{|\mu_i|}{2}}\mathbf{v}_i\right)^T\mathbf{x}\right)^2 - \sum_{\substack{i=1\\ \mu_i<0}}^{N}\left(\left(\sqrt{\frac{|\mu_i|}{2}}\mathbf{v}_i\right)^T\mathbf{x}\right)^2 + \mathbf{f}^T\mathbf{x} + c\,. \tag{9.6}$$

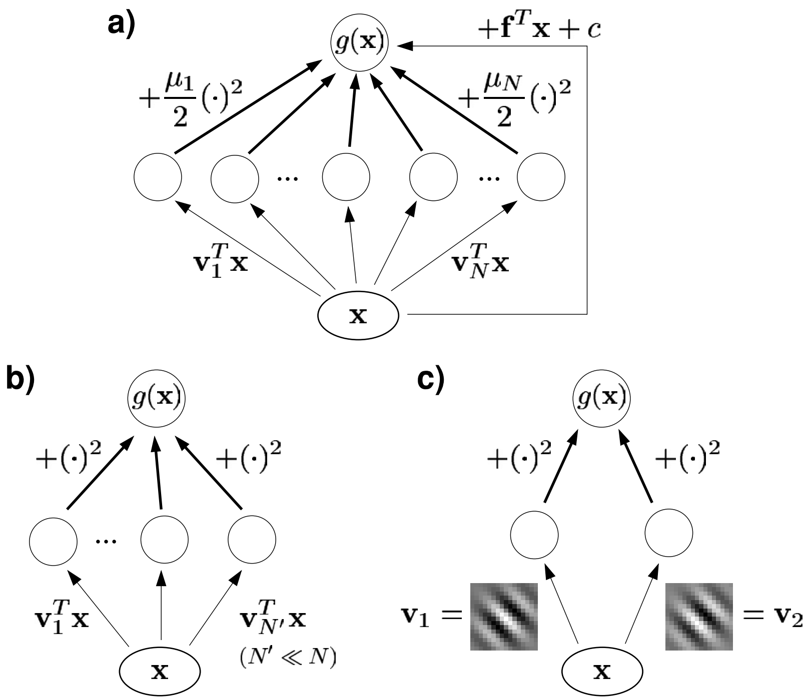

Figure 9: Neural networks related to inhomogeneous quadratic forms. In all plots we assume that the norm of the subunits is 1 (i.e., $\|\mathbf{v}_i\| = 1$). The ellipse in the input layer represents a multidimensional input. (a) Neural network equivalent to an inhomogeneous quadratic form. The first layer consists of N linear subunits, followed by a quadratic nonlinearity weighted by the coefficients $\mu_i/2$. The output neuron sums the contribution of each subunit plus the output of a direct linear connection from the input layer. (b) Simpler neural network used in some theoretical studies. The output of the linear subunits is squared but not weighted and can give only an excitatory (positive) contribution to the output. There is no direct linear connection between input and output layer. (c) The energy model of complex cells consists of two linear subunits whose weights are Gabor filters having the same shape except for a 90 degree phase difference. The output is given by the square sum of the response of the two subunits.

This equation also shows that the subunits are unique only up to an orthogonal transformation (i.e., a rotation or reflection) of the excitatory subunits and another one of the inhibitory subunits, which can be seen as follows. Let $\mathbf{A}^+$ and $\mathbf{A}^-$ be the matrices having as rows the vectors $\sqrt{|\mu_i|/2}\,\mathbf{v}_i$ for positive and negative μ_i, respectively. Equation 9.6 can then be

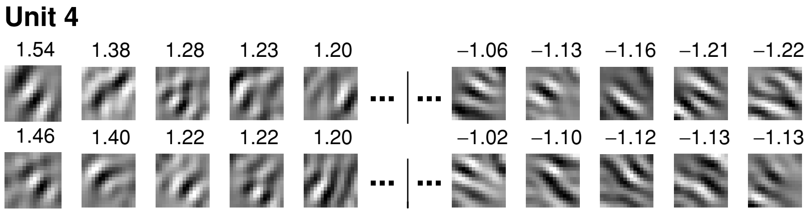

Figure 10: Random rotations of the positive and negative subunits. Two examples of the weights of the subunits of unit 4 after a random rotation as in equation 4.8. The numbers above the patches are the weighting coefficients on the second layer when the weight vectors of the first layer are normalized to norm 1. The subunits before rotation are equal to the eigenvectors of unit 4, and their weighting coefficients are equal to half the eigenvalues (see Figure 2, top).

rewritten as

$$g(\mathbf{x}) = \|\mathbf{A}^{+}\mathbf{x}\|^2 - \|\mathbf{A}^{-}\mathbf{x}\|^2 + \mathbf{f}^T\mathbf{x} + c\,. \tag{9.7}$$

Since the length of a vector does not change under rotation or reflection, the output of the function remains unchanged if we introduce two orthogonal transformations $\mathbf{R}^{+}$ and $\mathbf{R}^{-}$:

$$g(\mathbf{x}) = \|\mathbf{R}^{+}\mathbf{A}^{+}\mathbf{x}\|^2 - \|\mathbf{R}^{-}\mathbf{A}^{-}\mathbf{x}\|^2 + \mathbf{f}^T\mathbf{x} + c\,. \tag{9.8}$$

Figure 10 shows the weights of the subunits of the neural network equivalent to unit 4 as defined by the eigenvectors of $\mathbf{H}$ (see equation 9.4) after a random rotation of the excitatory and inhibitory subunits. The subunits are not as structured as in the case of the eigenvectors (cf. Figure 2), although the orientation and frequency can still be identified.

The neural model suggests alternative ways to learn quadratic forms, for example, by adapting the weights by backpropagation. The high number of parameters involved, however, could make it difficult for an incremental optimization method to avoid local extrema. On the other hand, each network of this form can be transformed into a quadratic form and analyzed with the techniques described in this article, which might be useful, for example, to compute the optimal stimuli and the invariances.

The equivalent neural network shows that quadratic forms are compatible with the hierarchical model of the visual cortex first proposed by Hubel and Wiesel (1962), in which complex cells pool over simple cells having similar orientation and frequency in order to achieve phase-shift invariance. This was later formalized in the energy model of complex cells (Adelson & Bergen, 1985), which can be implemented by the neural network introduced above. The subunits are interpreted as simple cells and the output unit as a complex cell. In its usual description, the energy model consists of only two excitatory subunits. If, for example, the subunits are two Gabor wavelets with identical envelope function, frequency, and orientation but with a 90 degree phase difference (see Figure 9c), the network will reproduce the basic properties of complex cells: edge detection and phase-shift invariance. Additional excitatory or inhibitory subunits might introduce additional complex cell invariances, broaden or sharpen the orientation and frequency tuning, and provide end or side inhibition. However, as mentioned in the previous section, the neural network is not unique, so that the subunits can assume different forms, many of which might not be similar to simple cells (see Figure 10). For example, as discussed in section 8, if the linear term is missing and the subunits are defined using the eigenvectors of **H** as in equation 9.4, the linear filters of the subunits can be interpreted as the optimal stimuli and the invariances thereof. As shown in Figure 2, the invariances themselves need not be structured like a simple cell, since they only represent transformations of the optimal stimuli.

10 Relation to Other Studies

As mentioned in section 1, quadratic forms occur in experimental studies as a second-order approximation of the receptive field of neurons. The linear and quadratic terms correspond in this case to the first two terms in a Wiener expansion. They can be estimated from a stimulus-response electrophysiological recording using the spike-triggered average (STA) and the spike-triggered covariance matrix (STC) (van Steveninck & Bialek, 1988; Lewis et al., 2002; Schwartz et al., 2002; Touryan et al., 2002; Rust et al., 2004; Simoncelli et al., 2004).

Most of these studies perform an analysis of the first principal components of the STC, which is motivated in terms of identifying the stimuli that contribute most to the variance of the output of the neuron (Lewis et al., 2002; Schwartz et al., 2002; Rust et al., 2004; Simoncelli et al., 2004) or more in terms of a gaussian approximation of the spike-triggered ensemble (van Steveninck and Bialek, 1988; Touryan et al., 2002). The extracted principal components span the subspace of stimuli that governs the response of a cell (Rust et al., 2004). If the linear term is negligible, our analysis is mostly consistent with this interpretation: ordering the eigenvectors by decreasing eigenvalues, the first one corresponds to the optimal stimulus and the following ones to the most relevant invariances (see section 8). Every stimulus that is generated by a linear combination of the optimal stimulus and the most relevant invariances is going to produce a strong output in the quadratic form. However, using the concept of invariances, we can refine the analysis and identify a hypercone in this subspace where the output is more than 80% of the maximal one with a large extension in the most invariant directions and a small one in the least invariant ones (see section 5).

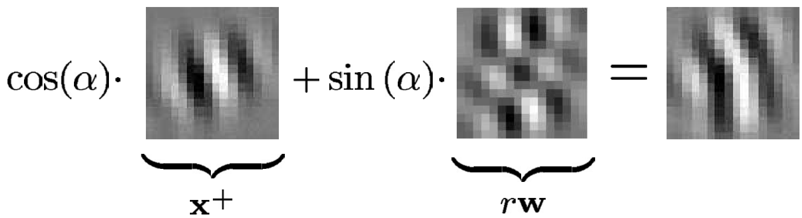

Figure 11: Interpretation of the invariances. This figure illustrates that although the vector corresponding to an invariance (center) might be difficult to interpret or even look unstructured, when applied to the optimal excitatory stimulus (left) it can code for a meaningful invariance (right). The invariance shown here is the curvature invariance of Figure 6f.

The stimuli lying in this hypercone are all nearly optimal stimuli, and their visualization can give good insight in the overall behavior of the neuron.

In our approach, the quadratic forms are interpreted as second-order approximations of the input-output functions computed by the neurons, and the resulting characterization is similar to the one given by classical physiological experiments (e.g., De Valois, Albrecht, & Thorell, 1982; De Valois, Yund, & Hepler, 1982; Schiller, Finlay, & Volman, 1976a, 1976b). Because of this interpretation, the linear term cannot be neglected or eliminated as in the experimental studies cited above. Only if the linear term is proved to be reasonably close to zero can one consider the quadratic term alone and apply the methods described in section 8.

Two recent theoretical studies (Hashimoto, 2003; Bartsch & Obermayer, 2003) learned quadratic forms without the linear term from natural images. The eigenvectors of $\mathbf{H}$ were visualized and interpreted as "relevant features." Some of them were discarded because they were "unstructured." According to our analysis, this interpretation holds for only the two eigenvectors with the largest positive and negative eigenvalues. We think that the remaining eigenvectors should not be visualized directly but applied as transformations to the optimal stimuli. Therefore, it is possible for them to look unstructured but still represent a structured invariance, as illustrated in Figure 11. For example, Hashimoto (2003, Fig. 5a) shows the eigenvectors of a quadratic form learned by a variant of SFA performed by gradient descent. The two largest eigenvectors look like two Gabor wavelets and have the same orientation and frequency. According to the interpretation above and to Hashimoto, this shows that the network responds best to an oriented stimulus and is invariant to a phase shift. The third eigenvector looks like a Gabor wavelet with the same frequency as the first two but a slightly different orientation. Hashimoto suggests that this eigenvector makes the interpretation of that particular quadratic form difficult. According to our analysis, that vector might code for a rotation invariance, which would be compatible with a complex cell behavior.

Neural networks closely related to those presented in section 9 were used in some theoretical studies (Hyvärinen & Hoyer, 2000, 2001; Körding et al., 2004). There, a small set of linear subunits (2 to 25) was connected to an output unit that took the sum of the squared activities (see Figure 9b). These networks differ from inhomogeneous quadratic forms in that they lack a direct linear contribution to the output and have much fewer subunits (a quadratic form of dimension N has N subunits). The most important difference, however, is related to the normalization of the weights. In the theoretical studies cited above, the weights are normalized to a fixed norm, and the activity of the subunits is not weighted. In particular, since there are no negative coefficients, no inhibition is possible, whereas this turned out to be essential for a number of complex cell properties in our simulations. However, the results of section 9 show that it is possible to use the algorithms presented here to analyze and interpret the weights of this kind of neural network.

11 Conclusion

We have presented a collection of tools to analyze nonlinear functions, in particular quadratic forms. These tools allow us to visualize the coefficients of the individual terms of an inhomogeneous quadratic form, to compute its optimal stimuli (i.e., the stimuli that maximize or minimize the quadratic form under a fixed energy constraint) and their invariances (i.e., the transformations of the optimal stimuli to which the quadratic form is most insensitive), and to determine which of these invariances are statistically significant. We have also proposed a way to measure the relative contribution of the linear and quadratic term. Moreover, we have discussed a neural network architecture equivalent to a given quadratic form. The methods presented here can be used in a variety of fields, in particular in physiological experiments to study the nonlinear receptive fields of neurons and in theoretical studies.

Acknowledgments

This work has been supported by a grant from the Volkswagen Foundation. We are grateful to Thomas Neukircher for some interesting discussions about some of the mathematical ideas of the article and to Henning Sprekeler and two anonymous reviewers for useful comments on the manuscript. The home page of the first author contains additional material to this article, including the animations corresponding to Figure 6 and Matlab source code for the algorithms of sections 4 and 5.

References

Adelson, E. H., & Bergen, J. R. (1985). Spatiotemporal energy models for the perception of motion. *Journal of the Optical Society of America A, 2*(2), 284–299.

Baddeley, R., Abbott, L., Booth, M., Sengpiel, F., Freeman, T., Wakeman, E., & Rolls, E. (1997). Responses of neurons in primary and inferior temporal visual cortices to natural scenes. *Proc. R. Soc. Lond. B Biol. Sci., 264*(1389), 1775–1783.

Bartsch, H., & Obermayer, K. (2003). Second-order statistics of natural images. *Neurocomputing, 52–54*, 467–472.

Berkes, P., & Wiskott, L. (2002). Applying slow feature analysis to image sequences yields a rich repertoire of complex cell properties. In J. R. Dorronsoro (Ed.), *Artificial Neural Networks—ICANN 2002 Proceedings* (pp. 81–86). Berlin: Springer.

Berkes, P., & Wiskott, L. (2005a). *On the analysis and interpretation of inhomogeneous quadratic forms as receptive fields*. Cognitive Sciences EPrint Archive (CogPrints) 4081, http://cogprints.org/4081/.

Berkes, P., & Wiskott, L. (2005b). Slow feature analysis yields a rich repertoire of complex cell properties. *Journal of Vision, 5*(6), 579–602, http://journalofvision.org/5/6/9, doi: 10.1167/5.6.9.

Dayan, P., & Abbott, L. F. (2001). *Theoretical neuroscience: Computational and mathematical modeling of neural systems*. Cambridge, MA: MIT Press.

De Valois, R., Albrecht, D., & Thorell, L. (1982). Spatial frequency selectivity of cells in macaque visual cortex. *Vision Res., 22*, 545–559.

De Valois, R., Yund, E., & Hepler, N. (1982). The orientation and direction selectivity of cells in macaque visual cortex. *Vision Res., 22*(5), 531–544.

Fortin, C. (2000). *A survey of the trust region subproblem within a semidefinite framework*. Unpublished master's thesis, University of Waterloo.

Hashimoto, W. (2003). Quadratic forms in natural images. *Network: Computation in Neural Systems, 14*(4), 765–788.

Hubel, D., & Wiesel, T. (1962). Receptive fields, binocular interaction and functional architecture in the cat's visual cortex. *Journal of Physiology, 160*, 106–154.

Hyvärinen, A., & Hoyer, P. (2000). Emergence of phase and shift invariant features by decomposition of natural images into independent feature subspaces. *Neural Computation, 12*(7), 1705–1720.

Hyvärinen, A., & Hoyer, P. (2001). A two-layer sparse coding model learns simple and complex cell receptive fields and topography from natural images. *Vision Research, 41*(18), 2413–2423.

Jones, J. P., & Palmer, L. A. (1987). An evaluation of the two-dimensional Gabor filter model of simple receptive fields in cat striate cortex. *Journal of Neurophysiology, 58*(6), 1233–1257.

Jutten, C., & Karhunen, J. (2003). Advances in nonlinear blind source separation. *Proc. of the 4th Int. Symp. on Independent Component Analysis and Blind Signal Separation (ICA2003)* (pp. 245–256).

Körding, K., Kayser, C., Einhäuser, W., & König, P. (2004). How are complex cell properties adapted to the statistics of natural scenes? *Journal of Neurophysiology, 91*(1), 206–212.

Lewis, E., Henry, K., & Yamada, W. (2002). Tuning and timing in the gerbil ear: Wiener-kernel analysis. *Hearing Research, 174*, 206–221.

MacKay, D. (1985). The significance of "feature sensitivity." In D. Rose & V. Dobson (Eds.), *Models of the visual cortex* (pp. 47–53). New York: Wiley.

Marmarelis, P., & Marmarelis, V. (1978). *Analysis of physiological systems: The white-noise approach*. New York: Plenum Press.

Pollen, D., & Ronner, S. (1981). Phase relationship between adjacent simple cells in the visual cortex. *Science, 212*, 1409–1411.

Rust, N. C., Schwartz, O., Movshon, J. A., & Simoncelli, E. (2004). Spike-triggered characterization of excitatory and suppressive stimulus dimensions in monkey V1. *Neurocomputing, 58–60*, 793–799.

Schiller, P., Finlay, B., & Volman, S. (1976a). Quantitative studies of single-cell properties in monkey striate cortex. I. Spatiotemporal organization of receptive fields. *J. Neurophysiol., 39*(6), 1288–1319.

Schiller, P., Finlay, B., & Volman, S. (1976b). Quantitative studies of single-cell properties in monkey striate cortex. II. Orientation specificity and ocular dominance. *J. Neurophysiol., 39*(6), 1320–1333.

Schölkopf, B., Smola, A., & Müller, K.-R. (1998). Nonlinear component analysis as a kernel eigenvalue problem. *Neural Computation, 10*(5), 1299–1319.

Schwartz, O., Chichilnisky, E., & Simoncelli, E. (2002). Characterizing neural gain control using spike-triggered covariance. In T. G. Dietterich, S. Becker, & Z. Ghahramani (Eds.), *Advances in neural information processing systems, 14* (pp. 269–276). Cambridge, MA: MIT Press.

Simoncelli, E. P., Pillow, J., Paninski, L., & Schwartz, O. (2004). Characterization of neural responses with stochastic stimuli. In M. Gazzaniga (Ed.), *The cognitive neurosciences* (3rd ed.). Cambridge, MA: MIT Press.

Stork, D., & Levinson, J. (1982). Receptive fields and the optimal stimulus. *Science, 216*, 204–205.

Touryan, J., Lau, B., & Dan, Y. (2002). Isolation of relevant visual features from random stimuli for cortical complex cells. *Journal of Neuroscience, 22*(24), 10811–10818.

van Steveninck, R., & Bialek, W. (1988). Coding and information transfer in short spike sequences. *Proc. Soc. Lond. B. Biol. Sci., 234*, 379–414.

Walter, W. (1995). *Analysis 2*. Berlin: Springer-Verlag.

Wiskott, L., & Sejnowski, T. (2002). Slow feature analysis: Unsupervised learning of invariances. *Neural Computation, 14*(4), 715–770.

Received February 16, 2005; accepted September 14, 2005.