The final version of this article has been published in
Neural Computation 15(9):2147-2177 (2003), published by The MIT Press.
This version does not differ significantly from the final version.

Slow Feature Analysis: A Theoretical Analysis of Optimal Free Responses

Laurenz Wiskott

Computational Neurobiology Laboratory
The Salk Institute for Biological Studies
San Diego, CA 92186-5800

Institute for Advanced Studies
Wallotstraße 19, D-14193 Berlin, Germany

Institute for Theoretical Biology*
Humboldt-University Berlin
Invalidenstraße 43, D-10115 Berlin, Germany
http://itb.biologie.hu-berlin.de/
l.wiskott@biologie.hu-berlin.de

Abstract

Temporal slowness is a learning principle that allows learning of invariant representations by extracting slowly varying features from quickly varying input signals. Slow feature analysis (SFA) is an efficient algorithm based on this principle, which has been applied to the learning of translation, scale, and other invariances in a simple model of the visual system. Here a theoretical analysis of the optimization problem solved by SFA is presented, which provides a deeper understanding of the simulation results obtained in previous studies.

1 Introduction

Temporal slowness as a learning principle is based on the observation that the environment, primary sensory signals, and internal representations of the environment change on different time scales. Our environment, e.g. the objects we see around us, changes usually on a slow time scale of several seconds. Primary sensory signals on the other hand, such as the responses of single receptors in the retina, change on a faster time scale, because even a small eye movement or shift of a textured object may lead to a rapid change of light intensity received by a receptor neuron. The internal representation of the environment, finally, should vary on a similar time scale as the environment itself, i.e. on a slow time scale. The sensory system does not have access to the environment but only to the primary sensory signal. The learning principle now assumes the following: If we succeed in extracting slowly varying features from the quickly varying sensory signal in a non-trivial way, then it is likely that we obtain a useful representation of the environment, which is in addition invariant or at least robust to frequent transformations of the sensory input, such as visual translation, scaling, rotation, or zoom.

*current address

This approach to unsupervised learning of invariant representations has been taken by a number of researchers since the early 90s (Földiák, 1991; Mitchison, 1991; Becker & Hinton, 1992; O'Reilly & Johnson, 1994; Stone & Bray, 1995; Wallis & Rolls, 1997; Peng et al., 1998; Kayser et al., 2001; Wiskott & Sejnowski, 2002) and an earlier description of the principle can be found in (Hinton, 1989, p. 208). Computational models based on the principle of temporal slowness have been quite successful in learning invariances in a number of contexts (see references above) and in reproducing receptive field properties of the primary visual cortex (Kayser et al., 2001; Berkes & Wiskott, 2002). However, there have been no attempts to also investigate the learning principle analytically in order to determine what kind of responses one might ideally expect from such a system. The paper presented here is a direct supplement to the paper (Wiskott & Sejnowski, 2002) and attempts to understand analytically some of the results that have been found numerically.

The paper is structured as follows: First the learning problem is stated in its full complexity as an optimization problem of variational calculus. Then a simplified optimization problem is derived that is more amenable to analytical treatment. A direct variational calculus approach for finding optimal solutions of the simplified optimization problem is given in Section 4. Section 5 presents an alternative algebraic approach. In Section 6 optimal responses are derived for a number of different boundary conditions. This includes a fairly detailed analysis of the results obtained in (Wiskott & Sejnowski, 2002, Examples 4, 5). The paper concludes with Section 7.

2 The Full Optimization Problem

The problem of extracting slow features from a quickly varying input signal can be formally stated as follows:

Optimization Problem 1 *Given an I-dimensional input signal $\mathbf{x}(t) = (x_1(t), ..., x_I(t))^T$ with time $t \in [t_A, t_B]$ and $(...)^T$ indicating the transpose. Find an input-output function $\mathbf{g}(\mathbf{x}) = (g_1(\mathbf{x}), ..., g_J(\mathbf{x}))^T$ generating the J-dimensional output signal $\mathbf{y}(t) = (y_1(t), ..., y_J(t))^T$ with $y_j(t) := g_j(\mathbf{x}(t))$ such that for each $j \in \{1, ..., J\}$*

$$\Delta_j := \Delta(y_j) := \langle \dot{y}_j^2 \rangle \quad \text{is minimal} \tag{1}$$

under the constraints

$$\begin{aligned} \langle y_j \rangle &= 0 \quad \text{(zero mean)}, & (2)\\ \langle y_j^2 \rangle &= 1 \quad \text{(unit variance)}, & (3)\\ \forall\, k < j: \quad \langle y_k\, y_j \rangle &= 0 \quad \text{(decorrelation)}, & (4) \end{aligned}$$

where the dot in $\dot{y}_j$ indicates the temporal derivative and angle brackets indicate temporal averaging, i.e. $\langle f \rangle := \frac{1}{t_B - t_A} \int_{t_A}^{t_B} f(t)\, dt$.

Equation (1) expresses the primary objective of temporal slowness by minimizing the temporal variation of the output signal. Constraints (2) and (3) help avoiding the trivial solution $y_j(t) = \text{const}$. Constraint (4) guarantees that different output signal components carry different information and do not simply reproduce each other. It also induces an order, so that $y_1(t)$ is the optimal output signal component, while $y_2(t)$ is a less optimal one, since it obeys the additional constraint $\langle y_1\, y_2 \rangle = 0$. Thus, $\Delta(y_k) \leq \Delta(y_j)$ if $k < j$.

3 The Simplified Optimization Problem

Optimization Problem 1 is too difficult to solve analytically in most practical cases. To simplify the problem we will now ignore the input signal and determine the optimal *free* output signal. The term *free* shall indicate the lack of constraints from an input signal or a class of input-output functions, but it permits constraints on the output signal itself, such as cyclic boundary conditions. Thereby we can investigate theoretically how the system would respond under idealized conditions. We formulate the simpler

Optimization Problem 2 *Find a J-dimensional output signal $\mathbf{y}(t) = (y_1(t), ..., y_J(t))^T$ with $t \in [t_A, t_B]$ such that for each $j \in \{1, ..., J\}$*

$$\Delta_j := \Delta(y_j) := \langle \dot{y}_j^2 \rangle \quad \textit{is minimal} \tag{5}$$

under the constraints

$$\begin{aligned} \langle y_j \rangle &= 0 \qquad \textit{(zero mean)}\,, &(6)\\ \langle y_j^2 \rangle &= 1 \qquad \textit{(unit variance)}\,, &(7)\\ \forall\, k < j: \quad \langle y_k\, y_j \rangle &= 0 \qquad \textit{(decorrelation)}\,, &(8) \end{aligned}$$

and possibly some boundary conditions, such as

$$\begin{aligned} y_j(t_A) &= y_{jA}\,, &(9)\\ y_j(t_B) &= y_{jB}\,. &(10) \end{aligned}$$

This problem can be analyzed with different methods. We will first apply standard variational calculus and then develop an alternative algebraic approach. In the following the term *response* will be used as a synonym for the term *output signal component* for brevity. In context of theoretical considerations responses are continuous functions, while in computer simulations they are discretely sampled.

4 Variational Calculus Approach

The direct and most general approach is variational calculus. From Objective (5) and Constraints (6–8) we derive the Lagrangian function for y_j:

$$L(t, y_j, \dot{y}_j, \lambda_{j0}, \lambda_{jj}, \lambda_{jk}) := \frac{1}{2}\dot{y}_j^2(t) + \lambda_{j0} y_j(t) + \lambda_{jj}\frac{1}{2} y_j^2(t) + \sum_{k<j} \lambda_{jk} y_j(t) y_k(t)\,, \tag{11}$$

where Objective (5) and Constraint (7) have been multiplied by a factor of $1/2$ for mathematical convenience without loss of generality. The corresponding Euler-Lagrange equation is

$$\frac{\partial}{\partial y_j} L - \frac{\mathrm{d}}{\mathrm{d}t}\frac{\partial}{\partial \dot{y}_j} L = \lambda_{j0} + \lambda_{jj} y_j(t) + \sum_{k<j} \lambda_{jk} y_k(t) - \ddot{y}_j(t) = 0\,. \tag{12}$$

Any solution of Optimization Problem 2 solves this differential equation. The free parameters λ_{j0}, λ_{jj} and λ_{jk} have to be chosen such that the constraints and boundary conditions are fulfilled. If no boundary conditions are given, they have to be varied to find the optimal solution. Notice that there may be solutions to (12), which are not solutions of Optimization Problem 2, because the Euler-Lagrange equation is only a necessary condition but not a sufficient one. If several solutions of the Euler-Lagrange equation exist, the optimal one has to be selected by additional considerations. The family of functions solving the Euler-Lagrange equation is given by

Theorem 1 *The optimal free responses for Optimization Problem 2 have the form*

$$y_j(t) = \sum_{k<j} d_{jk} y_k(t) + \begin{cases} -c_j/\lambda_{jj} + a_j \sin(\sqrt{-\lambda_{jj}}\, t) \; + b_j \cos(\sqrt{-\lambda_{jj}}\, t) & \text{if } \lambda_{jj} < 0 \\ c_j\, t^2/2 \;\; + a_j \qquad\qquad\qquad\qquad + b_j\, t & \text{if } \lambda_{jj} = 0 \\ -c_j/\lambda_{jj} + a_j \exp(+\sqrt{\lambda_{jj}}\, t) + b_j \exp(-\sqrt{\lambda_{jj}}\, t) & \text{if } \lambda_{jj} > 0 \end{cases} \tag{13}$$

with $c_j := (\lambda_{j0} - \sum_{k<j} d_{jk}\lambda_{k0})$ and d_{jk} chosen such that $(\lambda_{jj} d_{jk} + \lambda_{jk}) = \sum_{l<j} d_{jl}\lambda_{lk}$ for all $k < j$.

Proof First we notice that the Euler-Lagrange equation (12) is an inhomogeneous linear differential equation with constant coefficients. Thus its general solution y_j is the sum of a particular solution y_{pj} of the inhomogeneous equation and the general solution y_{gj} of the corresponding homogeneous equation, i.e. $y_j(t) = y_{pj}(t) + y_{gj}(t)$. Since the Euler-Lagrange equation is of second order, the general solution y_{gj} of the homogeneous equation is an arbitrary linear combination of two linearly independent functions. y_{gj} is easy to find and has the same form for all j:

$$y_{gj}(t) \quad := \quad \begin{cases} a_j \sin(\sqrt{-\lambda_{jj}}\, t) + b_j \cos(\sqrt{-\lambda_{jj}}\, t) & \text{if } \lambda_{jj} < 0 \\ a_j + b_j\, t & \text{if } \lambda_{jj} = 0 \\ a_j \exp(+\sqrt{\lambda_{jj}}\, t) + b_j \exp(-\sqrt{\lambda_{jj}}\, t) & \text{if } \lambda_{jj} > 0 \end{cases} \tag{14}$$

Which of these three types of solutions is the correct one depends on the boundary conditions $y_j(t_A) = y_{jA}$ and $y_j(t_B) = y_{jB}$. Numerical analysis of several examples indicate that in general the exponential solution is the correct one if one or both of the boundary values y_{jA} and y_{jB} are large in magnitude and that the oscillatory solution is the correct one if the boundary values are close to zero or can be optimized freely. If at least one boundary value y_{jA} or y_{jB} is large, the unit variance constraint requires the response to quickly go close to zero and stay there most of the time, which leads to the exponential solution. If none of the boundary values is constrained to be large, the oscillatory solution is preferable, because then the steep sections of the response, which are expensive in terms of the primary objective of slowness, are near zero, where they are most efficient in increasing the variance of the signal to fulfill the unit variance constraint. The linear solution marks the transition between these two cases. These are only intuitive arguments, of course, but since later we will use this variational calculus approach merely in a suggestive way to guess the right solutions, a more rigorous treatment of this issue is not necessary here. In this paper the boundary values are usually either close to zero or free to be optimized, so that the oscillatory solution is the common one.

The particular solutions y_{pj} can be derived by mathematical induction.

Part 1 (basis of induction): For $j = 1$ it is easy to show[1] that

$$y_{p1}(t) \quad := \quad \begin{cases} -\frac{\lambda_{10}}{\lambda_{11}} & \text{if } \lambda_{11} \neq 0 \\ \frac{\lambda_{10}}{2} t^2 & \text{if } \lambda_{11} = 0 \end{cases} \tag{15}$$

is a particular solution of the Euler-Lagrange equation (12).

[1] For $j = 1$ there is no $k < j$ and the Euler-Lagrange equation (12) for y_{p1} simplifies to

$$\lambda_{10} + \lambda_{11} y_{p1}(t) - \ddot{y}_{p1}(t) = 0\,.$$

Inserting y_{p1} and its second derivative

$$\ddot{y}_{p1}(t) \quad = \quad \begin{cases} 0 & \text{if } \lambda_{11} \neq 0 \\ \lambda_{10} & \text{if } \lambda_{11} = 0 \end{cases}$$

yields the true statements

$$\begin{aligned} \lambda_{10} + \lambda_{11}\left(-\frac{\lambda_{10}}{\lambda_{11}}\right) - 0 \quad &= \quad 0 \qquad \text{if } \lambda_{11} \neq 0\,, \\ \lambda_{10} + \underbrace{\lambda_{11}}_{=0} \frac{\lambda_{10}}{2} t^2 - \lambda_{10} \quad &= \quad 0 \qquad \text{if } \lambda_{11} = 0\,. \end{aligned}$$

Part 2 (inductive step): Assume Theorem 1 is true for all y_k with $k < j$. It can then be shown[2] that a particular solution of the Euler-Lagrange equation (12) for j is given by

$$y_{pj}(t) \quad := \quad \sum_{k<j} d_{jk} y_k(t) + \left\{ \begin{array}{ll} -c_j/\lambda_{jj} & \text{if } \lambda_{jj} \neq 0 \\ c_j\, t^2/2 & \text{if } \lambda_{jj} = 0 \end{array} \right. , \tag{16}$$

if the parameters c_j and d_{jk} are chosen according to Theorem 1. Adding the particular solution y_{pj} to the general solution y_{gj} defined by (14) yields y_j as defined by (13). ■

This variational calculus approach shows that the optimal free responses are typically oscillatory (or exponential, if the boundary values are large in magnitude). However, this kind of analysis is difficult for later responses with higher index j and more complex boundary conditions, e.g. if the solution has to be constant over a certain time interval. Thus in the next section we take a different approach that turns out to be simpler and more powerful.

5 Algebraic Approach

In the variational calculus approach the goal was to find optimal responses within an infinite-dimensional function space. In the algebraic approach we confine our analysis to a finite-dimensional space with dimensionality N. This is not a serious limitation, since in computer simulations the output signals have a finite dimensionality in any case and the dimensionality of our analysis can be arbitrarily high. Furthermore we assume that the responses are continuous functions (to exclude steps), piecewise differentiable and of L_2 (so that we may define an inner product of the form $\int a(t)b(t)\,\mathrm{d}t$). For simplicity we also assume that the responses we are looking for span the whole N-dimensional signal space, which implies $J = N$. This is again

[2] Inserting y_{pj} and its second derivative

$$\ddot{y}_{pj}(t) \quad = \quad \sum_{k<j} d_{jk} \ddot{y}_k(t) + \left\{ \begin{array}{ll} 0 & \text{if } \lambda_{jj} \neq 0 \\ c_j & \text{if } \lambda_{jj} = 0 \end{array} \right.$$

into the Euler-Lagrange equation (12) yields

$$\begin{aligned}
& \lambda_{j0} + \lambda_{jj} y_{pj}(t) + \sum_{k<j} \lambda_{jk} y_k(t) - \ddot{y}_{pj}(t) \\
= \; & \lambda_{j0} - c_j + \sum_{k<j} \left(\lambda_{jj} d_{jk} + \lambda_{jk}\right) y_k(t) - \sum_{k<j} d_{jk} \ddot{y}_k(t) \quad \text{(this holds for } \lambda_{jj} = 0 \text{ and } \lambda_{jj} \neq 0\text{)} \\
= \; & \sum_{k<j} d_{jk} \lambda_{k0} + \sum_{k<j} \left(\lambda_{jj} d_{jk} + \lambda_{jk}\right) y_k(t) - \sum_{k<j} d_{jk} \ddot{y}_k(t) \quad \text{(since } c_j := (\lambda_{j0} - \textstyle\sum_{k<j} d_{jk} \lambda_{k0})\text{)} \\
= \; & \sum_{k<j} d_{jk} \lambda_{k0} + \sum_{k<j} \sum_{l<j} d_{jl} \lambda_{lk} y_k(t) - \sum_{k<j} d_{jk} \ddot{y}_k(t) \\
& \quad \text{(since we chose } d_{jk} \text{ such that } \left(\lambda_{jj} d_{jk} + \lambda_{jk}\right) = \textstyle\sum_{l<j} d_{jl} \lambda_{lk} \text{ for all } k < j\text{)} \\
= \; & \sum_{k<j} d_{jk} \lambda_{k0} + \sum_{l<j} \sum_{k \leq l} d_{jl} \lambda_{lk} y_k(t) - \sum_{k<j} d_{jk} \ddot{y}_k(t) \quad \text{(since } \lambda_{lk} = 0 \text{ for all } l < k\text{)} \\
= \; & \sum_{k<j} d_{jk} \lambda_{k0} + \sum_{k<j} d_{jk} \sum_{l \leq k} \lambda_{kl} y_l(t) - \sum_{k<j} d_{jk} \ddot{y}_k(t) \\
= \; & \sum_{k<j} d_{jk} \underbrace{\left(\lambda_{k0} + \lambda_{kk} y_k(t) + \sum_{l<k} \lambda_{kl} y_l(t) - \ddot{y}_k(t) \right)}_{=0} \\
= \; & 0 \quad \text{(since each } y_k \text{ solves its corresponding Euler-Lagrange equation (12))}
\end{aligned}$$

To determine the $(J-1)$ free parameters d_{jk}, a linear system of $(J-1)$ equations must be solved. For such a system a solution always exists. If the equations are not all linearly independent, the solution may not be unique. Thus also y_{pj} may not be uniquely determined. However, this is not a problem since we need only one particular solution to the Euler-Lagrange equation (12).

no real limitation, since earlier responses are not affected in any way by later ones and we can always discard later ones, if we are not interested in that many responses. These restrictions permit a much more elegant and powerful analysis with algebraic methods following closely the logic of the SFA-Algorithm (Wiskott & Sejnowski, 2002).

Assume the J-dimensional space of responses we consider is given by a basis of linearly independent functions $a_1(t), ..., a_J(t)$ with zero mean. From such a basis we can always derive an orthonormal basis $b_1(t), ..., b_J(t)$ with the inner product defined by $\langle ab \rangle := \frac{1}{t_B - t_A} \int_{t_A}^{t_B} a(t)b(t)\,\mathrm{d}t$. Let $\mathbf{C}$ denote the covariance matrix with $C_{mn} := \langle b_m b_n \rangle$ and $\dot{\mathbf{C}}$ denote the matrix of the inner products of the time derivatives with $\dot{C}_{mn} := \langle \dot{b}_m \dot{b}_n \rangle$. $\dot{\mathbf{C}}$, like $\mathbf{C}$, has full rank, because the $\dot{b}_m$ are linearly independent, since they are derived from the linearly independent b_m by an invertible linear transformation. Notice also that $\dot{\mathbf{C}}$ is not a covariance matrix of the time derivatives, because the latter do not necessarily have zero mean. Since the functions b_m are orthonormal, they have not only zero mean, but also unit variance, and are mutually orthogonal, i.e. uncorrelated, so that $\mathbf{C} = \mathbf{1}$, with $\mathbf{1}$ indicating the unit matrix. The orientation of the orthonormal basis in space, however, is arbitrary.

Any valid response y_j is a linear combination of the basis functions, i.e. $y_j(t) = \sum_m w_{jm} b_m(t)$, with weight vector $\mathbf{w}_j = (w_{j1}, ..., w_{jJ})^T$. For a complete set of weight vectors $\mathbf{w}_j, j \in \{1, ..., J\}$ Optimization Problem 2 simplifies as follows:

$$\text{minimize} \quad \Delta_j = \langle \dot{y}_j^2 \rangle \quad = \quad \sum_{mn} w_{jm} \langle \dot{b}_m \dot{b}_n \rangle w_{jn} = \mathbf{w}_j^T \dot{\mathbf{C}} \mathbf{w}_j \tag{17}$$

under the constraints

$$\langle y_j \rangle \quad = \quad \sum_m w_{jm} \underbrace{\langle b_m \rangle}_{=0} = 0 \qquad \text{(zero mean)}\,, \tag{18}$$

$$\langle y_j^2 \rangle \quad = \quad \sum_{mn} w_{jm} \langle b_m b_n \rangle w_{jn} = \mathbf{w}_j^T \underbrace{\mathbf{C}}_{=\mathbf{1}} \mathbf{w}_j = \mathbf{w}_j^T \mathbf{w}_j = 1 \qquad \text{(unit variance)}\,, \tag{19}$$

$$\forall\, k < j : \quad \langle y_k\, y_j \rangle \quad = \quad \sum_{mn} w_{jm} \langle b_m b_n \rangle w_{kn} = \mathbf{w}_j^T \underbrace{\mathbf{C}}_{=\mathbf{1}} \mathbf{w}_k = \mathbf{w}_j^T \mathbf{w}_k = 0 \qquad \text{(decorrelation)}\,. \tag{20}$$

Constraint (18) is fulfilled automatically, since the basis functions have zero mean. Constraints (19) and (20) are fulfilled if and only if the weight vectors are orthonormal. Δ_1 is obviously minimal, if the (normalized) weight vector $\mathbf{w}_1$ is eigenvector of $\dot{\mathbf{C}}$ with smallest eigenvalue. $\mathbf{w}_2$ has to be chosen to correspond to the eigenvector with the second smallest eigenvalue (assuming non-degenerate eigenvalues), in order to yield a minimal value for Δ_2 under the decorrelation constraint (20). Similar arguments hold also for all other weight vectors, so that setting the weight vectors to the normalized eigenvectors of matrix $\dot{\mathbf{C}}$ ordered by increasing eigenvalue yields the (in general) unique solution of Optimization Problem 2 for a finite-dimensional space of responses. The Δ-values correspond to the eigenvalues, since

$$\Delta_j = \mathbf{w}_j^T \dot{\mathbf{C}} \mathbf{w}_j = \lambda_j \mathbf{w}_j^T \mathbf{w}_j = \lambda_j\,. \tag{21}$$

Notice that also the mixed inner products of the time derivatives of the responses vanish, which means that the time derivatives are mutually orthogonal, since

$$\forall\, k \neq j : \quad \langle \dot{y}_k\, \dot{y}_j \rangle \quad = \quad \sum_{mn} w_{jm} \langle \dot{b}_m \dot{b}_n \rangle w_{kn} = \mathbf{w}_j^T \dot{\mathbf{C}} \mathbf{w}_k = \lambda_k \mathbf{w}_j^T \mathbf{w}_k = 0\,. \tag{22}$$

This is a curious observation. Orthogonality of the time derivatives is even a sufficient criterion for a solution of Optimization Problem 2 under the given constraints, because only a set of eigenvectors of matrix $\dot{\mathbf{C}}$ yields functions with orthogonal time derivatives. As a consequence, any orthogonal set of functions with zero mean and unit variance for which also the time derivatives are mutually orthogonal forms a solution of Optimization Problem 2 within the space of responses spanned by these functions. This finding is so important that we state it as a theorem.

Theorem 2 *A set of functions y_j with the properties*

$$\langle y_j \rangle \quad = \quad 0 \qquad \textit{(zero mean)} \tag{23}$$

$$\begin{aligned}
\langle y_j^2 \rangle &= 1 \quad \textit{(unit variance)} && (24)\\
\langle y_j y_k \rangle &= 0 \quad \textit{(decorrelation)} && (25)\\
\langle \dot{y}_j \dot{y}_k \rangle &= 0 \quad \textit{(orthogonal time derivatives)} && (26)\\
j \le k &\Rightarrow \Delta_j \le \Delta_k \quad \textit{(order by slowness)} && (27)
\end{aligned}$$

is a solution of Optimization Problem 2 within the space Y spanned by these functions y_j. Such a function set is called Δ-optimal.

Remember that it is equivalent to say that two functions y_j and y_k are uncorrelated or that they are orthogonal, because the functions have zero mean. This is not true for the time derivatives, which must be orthogonal but not necessarily uncorrelated, because they may not have zero mean.

As mentioned above, the Δ-optimal set of functions y_j is unique (except for the signs) only if the eigenvalues of matrix $\dot{\mathbf{C}}$ are all different. If there are several orthonormal eigenvectors (weight vectors $\mathbf{w}_j$) with identical eigenvalues (Δ-values Δ_j), these eigenvectors define a subspace within which any other set of orthonormal vectors (weight vectors $\mathbf{w}'_j$) is an equally valid set of eigenvectors. Thus if we replace $\mathbf{w}_j$ by $\mathbf{w}'_j$ we obtain again a Δ-optimal set. Since the argument holds for any subset of weight vectors with identical Δ-value and since for a specific Δ-value the new weight vectors $\mathbf{w}'_j$ can be written as orthogonal linear combinations of the old weight vectors $\mathbf{w}_j$ we can state the following corollary.

Corollary 1 *Let $\{y_j \,|\, j = 1, ..., J\}$ be a Δ-optimal set of functions with Δ-values Δ_j. If $\mathbf{U}$ is an orthogonal $J \times J$ matrix with a block structure such that $U_{jp} = 0$ if $\Delta_j \neq \Delta_p$, then the transformed set of functions $\{y'_j \,|\, j = 1, ..., J\}$ with $y'_j(t) = \sum_p U_{jp} y_p(t)$ is also Δ-optimal with the same Δ-values as $\{y_j\}$.*

It is easy to prove this corollary in a direct fashion[3]. Another trivial but useful consequence of Theorem 2 is

Corollary 2 *Given two Δ-optimal function sets $\{y_j\}$ and $\{y'_k\}$ spanning the spaces Y and Y'. If all functions in Y are orthogonal to all functions in Y', i.e. $\langle y_j y'_k \rangle = 0 \; \forall j, k$, and the same holds true for the time derivatives, i.e. $\langle \dot{y}_j \dot{y}'_k \rangle = 0 \; \forall j, k$, then the union of $\{y_j\}$ and $\{y'_k\}$ ordered by its Δ-values forms a Δ-optimal set of the union of the spaces Y and Y'.*

This corollary justifies to consider odd and even functions separately, since these as well as their time derivatives are mutually orthogonal for symmetry reasons.

In the following we will consider some Δ-optimal sets of functions for different boundary conditions.

6 Optimal Free Responses

Before considering sets of Δ-optimal free responses, it is useful to introduce a measure of invariance that has a more intuitive interpretation than the Δ-value. We use here the index η (Wiskott & Sejnowski, 2002)

[3] We know that $\sum_p U_{jp} U_{kp} = \delta_{jk}$, since $\mathbf{U}$ is orthogonal, and $U_{jp}\Delta_p = U_{jp}\Delta_j$, since $U_{jp} = 0$ if $\Delta_j \neq \Delta_p$. Thus

$$\begin{aligned}
\langle y'_j \rangle &= \left\langle \sum_p U_{jp} y_p \right\rangle = \sum_p U_{jp} \underbrace{\langle y_p \rangle}_{=0} = 0\,,\\
\langle y'_j y'_k \rangle &= \left\langle \left(\sum_p U_{jp} y_p\right)\left(\sum_q U_{kq} y_q\right)\right\rangle = \sum_{pq} U_{jp} U_{kq} \underbrace{\langle y_p y_q \rangle}_{=\delta_{pq}} = \sum_p U_{jp} U_{kp} = \delta_{jk}\,,\\
\langle \dot{y}'_j \dot{y}'_k \rangle &= \left\langle \left(\sum_p U_{jp} \dot{y}_p\right)\left(\sum_q U_{kq} \dot{y}_q\right)\right\rangle = \sum_{pq} U_{jp} U_{kq} \underbrace{\langle \dot{y}_p \dot{y}_q \rangle}_{=\delta_{pq}\Delta_p} = \sum_p U_{jp} U_{kp} \Delta_p\\
&= \sum_p U_{jp} U_{kp} \Delta_j = \Delta_j \underbrace{\sum_p U_{jp} U_{kp}}_{\delta_{jk}} = \Delta_j \delta_{jk}\,,
\end{aligned}$$

and the set $\{y'_j\}$ is Δ-optimal according to Theorem 2 with the same Δ-values as $\{y_j\}$.

defined by

$$\eta(y) := \frac{D}{2\pi}\sqrt{\Delta(y)} \tag{28}$$

if $t \in [t_A, t_A + D]$. For a pure sine wave $y(t) := \sqrt{2}\sin(n\,2\pi\,t/D)$ with an integer number of oscillations n the index $\eta(y)$ is just the number of oscillations, i.e. $\eta = n$. Thus the index η of an arbitrary signal indicates what the number of oscillations would be for a pure sine wave of same Δ-value, at least for integer values of η. We also define $\eta_j := \eta(y_j)$.

We will now consider some examples of Δ-optimal sets of responses. The considerations are mainly based on Theorem 2.

6.1 Cyclic Boundary Condition

What is the Δ-optimal set of responses with cyclic boundary condition on the interval $[t_A, t_B]$? We know from Fourier analysis, that any continuous function with cyclic boundary condition on this interval can be written as a sum of sine and cosine functions $\sin(n2\pi\frac{t-t_A}{D})$ and $\cos(n2\pi\frac{t-t_A}{D})$ with integers n and $D := t_B - t_A$. We also know that these functions as well as their time derivatives are mutually orthogonal (which does not mean that the functions are orthogonal to the time derivatives). Thus the set of all sine and cosine functions up to a maximum frequency,

$$y_j(t) = \begin{cases} \sqrt{2}\sin\left((j+1)\pi\,\frac{t-t_A}{D}\right) & \text{if } j \text{ odd} \\ \sqrt{2}\cos\left(j\pi\,\frac{t-t_A}{D}\right) & \text{if } j \text{ even} \end{cases} \qquad t \in [t_A, t_B]\,, \tag{29}$$

$$\eta_j = \begin{cases} (j+1)/2 & \text{if } j \text{ odd} \\ j/2 & \text{if } j \text{ even} \end{cases}\,, \tag{30}$$

forms a Δ-optimal set, with $(j+1)/2$ and $j/2$ full oscillations for odd and even j, respectively, resulting in the corresponding η-values; see Figure 1. This set it not unique, however, since successive pairs of functions have identical η-values.

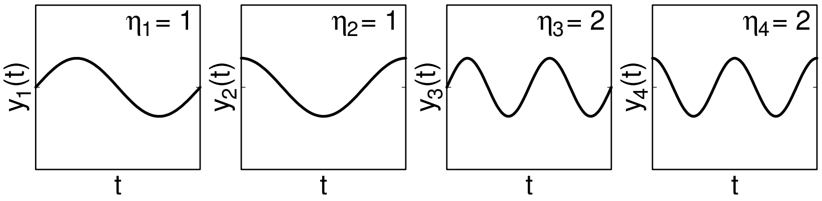

Figure 1: First four Δ-optimal responses for the cyclic boundary condition. t-axes range from t_A to t_B; y-axes range from -4 to $+4$.

6.2 Free Boundary Conditions

Consider now the more general case of all continuous functions on the interval $[t_A, t_B]$ without any further boundary condition. What is the corresponding Δ-optimal set? Assume $[t_A, t_B] = [0, \pi]$ for simplicity and without loss of generality. Any function with free boundary condition on the interval $[0, \pi]$ can be considered one half of a corresponding even function on the interval $[-\pi, \pi]$. Since we know from Fourier analysis that the cosine functions $\cos(nt)$ with integers n span the space of even functions on $[-\pi, \pi]$, they also span the space of any functions on $[0, \pi]$. For symmetry reasons $\int_{-\pi}^{0}\cos(j\,t)\cos(k\,t)\,\mathrm{d}t = \int_{0}^{+\pi}\cos(j\,t)\cos(k\,t)\,\mathrm{d}t$. Thus, since these functions are orthogonal on the interval $[-\pi, \pi]$, they are also orthogonal on the interval $[0, \pi]$. Similar arguments hold for the time derivatives. Generalizing these considerations to the interval $[t_A, t_B]$ leads to the Δ-optimal set

$$y_j(t) = \sqrt{2}\cos\left(j\pi\,\frac{t-t_A}{D}\right) \qquad t \in [t_A, t_B]\,, \tag{31}$$

$$\eta_j = j/2\,. \tag{32}$$

This set is unique (except for the signs), since each function has a different η-value. The first four functions are shown in Figure 2.

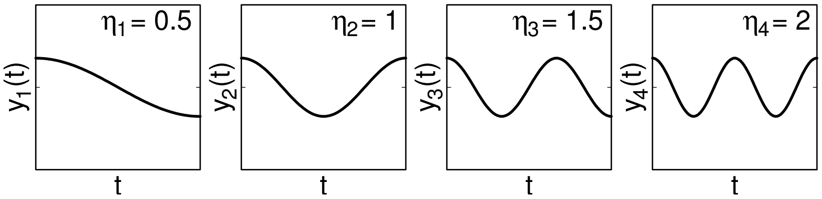

Figure 2: First four Δ-optimal responses for the free boundary condition. t-axes range from t_A to t_B; y-axes range from -4 to $+4$.

6.3 One Pattern

After the simple cases of cyclic and free boundary conditions, which give a first intuition of the nature of typical Δ-optimal sets, we will now derive results for the Examples 4 and 5 in (Wiskott & Sejnowski, 2002). In these examples a hierarchical network performing SFA was considered as a simple model of the visual system. The network had a one-dimensional retina as an input layer and nine units in the output layer, extracting the first nine responses y_j. The network was trained with several patterns that were presented one by one to the retina for a certain amount of time and with intermissions of no pattern presentation in between. The patterns were either moved translationally across the retina or they changed according to some other transformation (scale, 1D-rotation angle, contrast, illumination) or a combination of them. Here we relate only to simulation results obtained for translation and scale invariance; cf. (Wiskott & Sejnowski, 2002, Figs. 11, 21) and Fig. 10. A boundary condition that this training schedule imposes on the output signal and that we can consider in our current analysis is that the responses have the same constant values during all the time intervals where no pattern is presented to the network. In case of size invariance, due to the symmetry of the training with patterns increasing and decreasing in size, there is the additional constraint that the responses to single patterns must be even. Thus, odd responses must be disregarded in a comparison with the size invariance simulations of (Wiskott & Sejnowski, 2002).

First consider the simplest case of one pattern presentation. Let $[t_A, t_B]$ be the total time interval considered and $[t_a, t_b] \subset [t_A, t_B]$ the shorter time interval during which the pattern is presented to the network. The boundary condition requires $y_j(t) = c_j, \forall t \in [t_A, t_B] \setminus [t_a, t_b]$ with suitable constants c_j. What is the corresponding Δ-optimal set of responses?

Consider first an approximation by taking the limit $(t_B - t_A) \to \infty$. In that case $c_j \to 0$ due to the zero mean constraint and any average value of y_j within the interval $[t_a, t_b]$ can be compensated for by an infinitesimally small value of c_j. Assuming $c_j = 0$ and without the need to respect the zero mean constraint within the interval $[t_a, t_b]$ one can guess in analogy to the previous examples that a Δ-optimal set is approximately given by

$$y_j(t) = \begin{cases} \sqrt{2D/d}\,\sin\left(j\pi\frac{t-t_a}{d}\right) & \text{if } t \in [t_a, t_b] \\ 0 & \text{otherwise} \end{cases}, \tag{33}$$

$$\eta_j = jD/(2d)\,, \tag{34}$$

with $D := t_B - t_A$ and $d := t_b - t_a$. The first four responses of this set are shown in Figure 3.

For an intuitive understanding of this set assume, without loss of generality, $t_a = -\pi$ and $t_b = +\pi$. The functions with even index j can then be written[4] as $y_j(t) = \sqrt{D/\pi}\,(-1)^{\frac{j}{2}} \sin\left(\frac{j}{2}t\right)$. These are full sine waves with an integer number of oscillations within the interval $[-\pi, +\pi]$. Thus they are odd functions and therefore have zero mean exactly even for finite D, they are mutually orthogonal, and their time derivatives

[4] For even j

$$y_j(t) = \sqrt{2D/(2\pi)}\,\sin\left(j\pi\frac{t-(-\pi)}{2\pi}\right) = \sqrt{D/\pi}\,\sin\left(\frac{j}{2}t + \frac{j}{2}\pi\right) = \sqrt{D/\pi}\,(-1)^{\frac{j}{2}}\sin\left(\frac{j}{2}t\right).$$

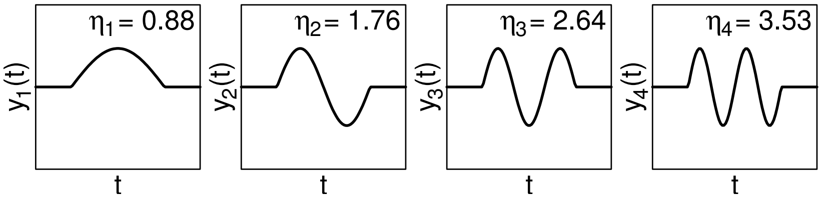

Figure 3: First four approximate Δ-optimal responses for the single pattern boundary condition. $D = 150$ and $d = 85$; t-axes range from t_A to t_B; y-axes range from -4 to $+4$. y_1 and y_3 do not have zero mean, while y_2 and y_4 have. Notice also that η_1 has an unrealistically low value, because the lowest possible η-value under the weaker cyclic boundary condition is 1; cf. (30) and Fig. 1.

are mutually orthogonal, too. Hence, this set of odd functions (with even index) forms a Δ-optimal set and from Fourier analysis it is known that they span the space of all continuous odd functions with boundary condition $y(-\pi) = y(+\pi) = 0$ (up to a certain frequency if j is limited).

The functions with odd index j can be written[5] as $y_j(t) = \sqrt{D/\pi}\,(-1)^{\frac{j-1}{2}} \cos\left(\frac{j}{2}t\right)$. These are cosine waves with an integer number of oscillations minus half an oscillation within the interval $[-\pi, +\pi]$. Thus they are even functions and, for symmetry reasons, they as well as their time derivatives are orthogonal to each other (following an argumentation similar to that in Section 6.2; a mathematical proof is given below). However, their mean value does not vanish for finite D, which means that they fulfill the requirements of a Δ-optimal set exactly only in the limit $D \to \infty$ and approximately for finite D. From Fourier analysis can be inferred that they span the space of all continuous even functions with boundary condition $y(-\pi) = y(+\pi) = 0$ (up to a certain frequency if j is limited). This can be seen as follows. We know from Section 6.2 that the set $\{\cos\left(\frac{j}{2}t\right)\}$ spans the space of all continuous functions with free boundary conditions on the interval $[0, 2\pi]$. Taking only the functions with odd index value yields a set that spans the space of all functions with an odd symmetry with respect to the reference point $+\pi$ and free boundary conditions on the interval $[0, 2\pi]$. For symmetry reasons this same set spans also the space of all functions with boundary condition $y(+\pi) = 0$ on the interval $[0, +\pi]$ and the space of all even functions (with reference point 0) with boundary condition $y(-\pi) = y(+\pi) = 0$ on the interval $[-\pi, +\pi]$. The latter is the required result.

Taking together odd and even functions of (33) they span the space of all continuous functions with

[5] For odd j

$$y_j(t) = \sqrt{2D/(2\pi)}\,\sin\left(j\pi\frac{t-(-\pi)}{2\pi}\right) = \sqrt{D/\pi}\,\sin\left(\frac{j}{2}t + \frac{j-1}{2}\pi + \frac{1}{2}\pi\right) = \sqrt{D/\pi}\,(-1)^{\frac{j-1}{2}} \cos\left(\frac{j}{2}t\right).$$

boundary condition $y(-\pi) = y(+\pi) = 0$ on the interval $[-\pi, +\pi]$. It can also be shown[6] more formally that (33) defines a Δ-optimal set in the limit $D \to \infty$ and that it fulfills the conditions for a Δ-optimal set approximately for finite D. Unfortunately, the latter is only suggestive and does not necessarily imply that the given y_j are an approximation of the true Δ-optimal set for finite D. However, it can be verified numerically that this is indeed the case. Notice also that the conditions for a Δ-optimal set are fulfilled exactly by the set of odd functions y_j (with even index j) because they have zero mean within the interval $[t_a, t_b]$, which means that at least they form a Δ-optimal set for finite D, but they span only part of the interesting function space.

To determine at least the first Δ-optimal even function exactly for the single pattern boundary condition we can take the variational calculus approach. To simplify the analysis and without loss of generality assume $t_A = -t_B$ and $t_a = -t_b$. We infer from Section 4 that the first Δ-optimal even function is of the form $y_1(t) = -\lambda_{10}/\lambda_{11} + b_1 \cos(\sqrt{-\lambda_{11}}\,t)$, since (i) the solution with $\lambda_{11} < 0$ is selected because the boundary value c_1 can be optimized freely, (ii) there are no terms with constants d_{1k} because no $k < 1$ exist, and (iii) $a_1 = 0$ because the solution has to be even. The constants $\lambda_{10}, \lambda_{11}$, and b_1 have to be chosen such that y_1 fulfills the constraints (6) and (7) and optimizes the Δ-value (5). I have found numerically with standard optimization techniques that the optimal λ_{10} is close to zero for all valid values of t_B, t_A, t_b, and t_a. Assuming vanishing λ_{10}, setting $a_1 = b_1$ and $\omega = \sqrt{-\lambda_{11}}$, generalizing to arbitrary values of t_a and t_b,

[6] With (33) we can verify that

$$\begin{aligned}
\langle y_j \rangle &= \frac{1}{D}\int_{t_A}^{t_B} y_j(t)\,\mathrm{d}t = \frac{\sqrt{2D/d}}{D}\int_{t_a}^{t_b} \sin\left(j\pi\frac{t-t_a}{d}\right)\mathrm{d}t = \sqrt{\frac{2}{Dd}}\int_{0}^{j\pi} \sin(t')\,\frac{d}{j\pi}\mathrm{d}t' \\
&= \sqrt{\frac{2d}{D}}\,\frac{1}{j\pi}\,(1-\cos(j\pi)) = \begin{cases} \sqrt{\frac{2d}{D}}\,\frac{2}{j\pi} & \text{for odd } j \\ 0 & \text{for even } j \end{cases} \\
\Longrightarrow \quad \lim_{D\to\infty} \langle y_j \rangle &= 0\,, \\
\langle y_j y_k \rangle &= \frac{1}{D}\int_{t_A}^{t_B} y_j(t) y_k(t)\,\mathrm{d}t = \frac{2D/d}{D}\int_{t_a}^{t_b} \sin\left(j\pi\frac{t-t_a}{d}\right)\sin\left(k\pi\frac{t-t_a}{d}\right)\mathrm{d}t \\
&= \frac{2}{d}\int_{0}^{\pi} \sin(jt')\sin(kt')\,\frac{d}{\pi}\mathrm{d}t' = \frac{1}{\pi}\int_{-\pi}^{+\pi} \sin(jt')\sin(kt')\,\mathrm{d}t' = \begin{cases} 0 & \text{if } j \neq k \\ 1 & \text{if } j = k \end{cases} .
\end{aligned}$$

The second last step is valid, because $\sin(jt')$ and $\sin(kt')$ are odd functions and the product $\sin(jt')\sin(kt')$ therefore an even function. Thus the integral over $[0, +\pi]$ can be replaced by the integral over $[-\pi, +\pi]$ divided by 2. The last step is valid, because sine waves with different but integer numbers of oscillations are orthogonal (case $j \neq k$) and because $\int_{-\pi}^{+\pi} \sin^2(jt')\,\mathrm{d}t' = \pi$ (case $j = k$).

For the time derivatives $\dot{y}_j$ we find similarly

$$\begin{aligned}
\langle \dot{y}_j \dot{y}_k \rangle &= \frac{1}{D}\int_{t_A}^{t_B} \dot{y}_j(t)\dot{y}_k(t)\,\mathrm{d}t = \frac{2D/d}{D}\,\frac{jk\pi^2}{d^2}\int_{t_a}^{t_b} \cos\left(j\pi\frac{t-t_a}{d}\right)\cos\left(k\pi\frac{t-t_a}{d}\right)\mathrm{d}t \\
&= \frac{2jk\pi^2}{d^3}\int_{0}^{\pi} \cos(jt')\cos(kt')\,\frac{d}{\pi}\mathrm{d}t' = \frac{jk\pi}{d^2}\int_{-\pi}^{+\pi} \cos(jt')\cos(kt')\,\mathrm{d}t' = \begin{cases} 0 & \text{for } j \neq k \\ \frac{j^2\pi^2}{d^2} & \text{for } j = k \end{cases} \\
\eta_j &= \frac{D}{2\pi}\sqrt{\langle \dot{y}_j^2 \rangle} = \frac{Dj}{2d}\,.
\end{aligned}$$

Notice that the conditions $\langle y_j y_k \rangle = \delta_{jk}$ and $\langle \dot{y}_j \dot{y}_k \rangle = \delta_{jk}\Delta_j$ are fulfilled exactly even for finite D.

and taking into account the zero mean and unit variance constraints, we have[7] for $j = 1$

$$
\begin{aligned}
y_j(t) &= \left\{ \begin{array}{cl} a_j \cos\left(\omega_j \frac{2t-(t_a+t_b)}{d}\right) & \text{if } t \in [t_a, t_b] \\ a_j \cos(\omega_j) & \text{otherwise} \end{array} \right. , && (35) \\
\text{with} \quad \frac{\tan(\omega_j)}{\omega_j} &= -(D/d-1) \quad \text{and} \quad \omega_j \in (j\pi/2, (j+1)\pi/2]\,, && (36) \\
a_j &= \sqrt{2/\Big(1+\big(1-d/D\big)\cos(2\omega_j)+(d/D)\sin(2\omega_j)/(2\omega_j)\Big)}\,, && (37) \\
\eta_j &= \frac{a_j}{2\pi}\sqrt{\omega_j(D/d)\big(2\omega_j-\sin(2\omega_j)\big)}\,. && (38)
\end{aligned}
$$

If D and d are given, the optimal frequency ω_j can be determined with (36) and the optimal amplitude with (37). (36) defines ω_j only implicitly. Thus it is helpful to draw the graphs of $(D/d-1)\omega_j$ and $-\tan(\omega_j)$ in a common diagram and take their first non-zero intersection as the solution. Such a graph is shown in Figure 4. It illustrates that ω_j lies between $\pi/2$ and π. The intersection between 1.5π and 2π is suboptimal, since it results in a larger η-value (38). (It is also an illustrative exercise to consider ω_j as given and determine D as a function of d with (36).)

[7] Assuming $t_a = -t_b$ for simplicity and without loss of generality and dropping index $j = 1$ for notational convenience we have

$$
\begin{aligned}
&& y(t) &= \left\{ \begin{array}{cl} a\cos\left(\omega\frac{2t}{d}\right) & \text{if } t \in [-t_b, t_b] \\ a\cos(\omega) & \text{otherwise} \end{array} \right. \qquad (\text{with } \omega \in (\pi/2, \pi] \text{ and } a > 0) \\
0 &\overset{!}{=}& \langle y \rangle &= \left((D-d)a\cos(\omega) \,+\, a\underbrace{\int_{-t_b}^{t_b} \cos\left(\omega\frac{2t}{d}\right)\mathrm{d}t}_{d\sin(\omega)/\omega}\right)\Big/D \\
&\Longleftrightarrow& \frac{\tan(\omega)}{\omega} &= -(D/d-1) \qquad (\text{since } \cos(\omega) \neq 0) \\
1 &\overset{!}{=}& \langle y^2 \rangle &= \left((D-d)a^2 \underbrace{\cos(\omega)^2}_{(1+\cos(2\omega))/2} + a^2 \underbrace{\int_{-t_b}^{t_b} \cos\left(\omega\frac{2t}{d}\right)^2 \mathrm{d}t}_{(d+d\sin(2\omega)/(2\omega))/2}\right)\Big/D \\
&& &= \Big((D-d)a^2/2 \,+\, (D-d)a^2\cos(2\omega)/2 \,+\, a^2 d/2 \,+\, a^2 d\sin(2\omega)/(2\omega)/2\Big)/D \\
&& &= a^2\Big(1+(1-d/D)\cos(2\omega)+(d/D)\sin(2\omega)/(2\omega)\Big)/2 \\
&\Longleftrightarrow& a &= \sqrt{2/\Big(1+(1-d/D)\cos(2\omega)+(d/D)\sin(2\omega)/(2\omega)\Big)} \qquad (\text{since } a > 0) \\
&& \langle \dot{y}^2 \rangle &= a^2\left(\omega\frac{2}{d}\right)^2 \underbrace{\int_{-t_b}^{t_b} \sin\left(\omega\frac{2t}{d}\right)^2 \mathrm{d}t}_{(d-d\sin(2\omega)/(2\omega))/2} \Big/D \quad = \quad a^2\omega(1/dD)(2\omega-\sin(2\omega)) \\
&\Longleftrightarrow& \eta &= \frac{D}{2\pi}\sqrt{\langle \dot{y}^2 \rangle} \quad = \quad \frac{a}{2\pi}\sqrt{\omega(D/d)(2\omega-\sin(2\omega))}
\end{aligned}
$$

Interestingly, it can be shown[8] that the solutions belonging to different intersections are mutually orthogonal and that also their time derivatives are orthogonal, so that they fulfill the conditions of Theorem 2 and therefore form a Δ-optimal set of responses. Thus Equations (35–38) can be taken for all odd indices j up to some arbitrary limit (the exact responses for even index j are still given by Eqs. (33, 34)). However, it is not as clear as in the previous sections that these responses actually span the space of all continuous functions up to a certain frequency, because the term frequency is not well defined here and we cannot resort to Fourier theory that easily.

To investigate this issue further consider the limiting cases $D = d$ and $D \to \infty$ for odd j. For $D = d$ we find $\omega_j = (j+1)\pi/2$ and $a_j = \sqrt{2}$ so that y_j are equal to $y_{(j+1)}$ for the cyclic boundary condition in Section 6.1. If we let D go to infinity we obtain $\lim_{D\to\infty} \omega_j = j\pi/2$. Taking the analytical limes of a_j is difficult since ω_j is given only implicitly, but it is intuitively clear and can be confirmed numerically, that a_j grows to large values. In this latter case the y_j become equal to their approximate counterparts of Equation (33). This also holds for finite D if j goes to infinity.

Thus, in the limiting cases $D = d$ and $D \to \infty$ (or $j \to \infty$) the responses given by Equations (35–37) converge to complete sets of all even functions up to a certain frequency in the sense of Fourier theory. This at least suggests that the exact Δ-optimal set for one pattern (Eq. (33) for even index and Eqs. (35–37) for odd index j) is also complete up to a certain frequency. The first four exact Δ-optimal responses for the

[8] First we show that the inner product between y_j and y_k vanishes, again assuming $t_a = -t_b$ for simplicity and without loss of generality:

$$
\begin{aligned}
0 \overset{?}{=} \langle y_j y_k \rangle &= \left((D-d) a_j \cos(\omega_j) a_k \cos(\omega_k) + a_j a_k \int_{-tr}^{t_b} \cos\left(\omega_j \frac{2t}{d}\right) \cos\left(\omega_k \frac{2t}{d}\right) \mathrm{d}t\right)\Big/ D \\
\iff \quad 0 &= (D-d)\cos(\omega_j)\cos(\omega_k) + \int_{-tr}^{t_b} \frac{1}{2}\left(\cos\left((\omega_j - \omega_k)\frac{2t}{d}\right) + \cos\left((\omega_j + \omega_k)\frac{2t}{d}\right)\right) \mathrm{d}t \\
&= (D-d)\cos(\omega_j)\cos(\omega_k) + \frac{d\sin(\omega_j - \omega_k)}{2(\omega_j - \omega_k)} + \frac{d\sin(\omega_j + \omega_k)}{2(\omega_j + \omega_k)} \qquad (\text{since } \omega_j \neq \omega_k) \\
\iff \quad 0 &= (\omega_j - \omega_k)(\omega_j + \omega_k)(D/d - 1)\cos(\omega_j)\cos(\omega_k) + (\omega_j + \omega_k)\sin(\omega_j - \omega_k)/2 \\
&\quad + (\omega_j - \omega_k)\sin(\omega_j + \omega_k)/2 \\
&= (\omega_j^2 - \omega_k^2)(D/d - 1)\cos(\omega_j)\cos(\omega_k) + (\omega_j + \omega_k)(\sin(\omega_j)\cos(\omega_k) - \cos(\omega_j)\sin(\omega_k))/2 \\
&\quad + (\omega_j - \omega_k)(\sin(\omega_j)\cos(\omega_k) + \cos(\omega_j)\sin(\omega_k))/2 \\
&= (\omega_j^2 - \omega_k^2)(D/d - 1)\cos(\omega_j)\cos(\omega_k) + \omega_j \sin(\omega_j)\cos(\omega_k) - \omega_k \cos(\omega_j)\sin(\omega_k) \\
\iff \quad 0 &= \frac{(\omega_j^2 - \omega_k^2)}{\omega_j^2 \omega_k^2}(D/d - 1) + \frac{1}{\omega_k^2}\frac{\sin(\omega_j)}{\omega_j \cos(\omega_j)} - \frac{1}{\omega_j^2}\frac{\sin(\omega_k)}{\omega_k \cos(\omega_k)} \qquad (\text{since } \omega_j^2 \omega_k^2 \cos(\omega_j)\cos(\omega_k) \neq 0) \\
&= \left(\frac{1}{\omega_k^2} - \frac{1}{\omega_j^2}\right)(D/d - 1) - \frac{1}{\omega_k^2}(D/d - 1) + \frac{1}{\omega_j^2}(D/d - 1) \qquad \left(\text{since } \tfrac{\tan(\omega_j)}{\omega_j} = -(D/d - 1)\ (36)\right) \\
&= 0
\end{aligned}
$$

Then we show that also the time derivatives are orthogonal:

$$
\begin{aligned}
0 \overset{?}{=} \langle \dot{y}_j \dot{y}_k \rangle &= a_j a_k \left(\omega_j \frac{2}{d}\right)\left(\omega_k \frac{2}{d}\right) \int_{-tr}^{t_b} \sin\left(\omega_j \frac{2t}{d}\right) \sin\left(\omega_k \frac{2t}{d}\right) \mathrm{d}t \Big/ D \\
\iff \quad 0 &= \int_{-tr}^{t_b} \frac{1}{2}\left(\cos\left((\omega_j - \omega_k)\frac{2t}{d}\right) - \cos\left((\omega_j + \omega_k)\frac{2t}{d}\right)\right) \mathrm{d}t \\
&= \frac{d\sin(\omega_j - \omega_k)}{2(\omega_j - \omega_k)} - \frac{d\sin(\omega_j + \omega_k)}{2(\omega_j + \omega_k)} \qquad (\text{since } \omega_j \neq \omega_k) \\
\iff \quad 0 &= (\omega_j + \omega_k)\sin(\omega_j - \omega_k)/2 - (\omega_j - \omega_k)\sin(\omega_j + \omega_k)/2 \\
&= (\omega_j + \omega_k)\Big(\sin(\omega_j)\cos(\omega_k) - \cos(\omega_j)\sin(\omega_k)\Big)/2 - (\omega_j - \omega_k)\Big(\sin(\omega_j)\cos(\omega_k) + \cos(\omega_j)\sin(\omega_k)\Big)/2 \\
&= \omega_k \sin(\omega_j)\cos(\omega_k) - \omega_j \cos(\omega_j)\sin(\omega_k) \\
\iff \quad 0 &= \frac{\sin(\omega_j)}{\omega_j \cos(\omega_j)} - \frac{\sin(\omega_k)}{\omega_k \cos(\omega_k)} \qquad (\text{since } \omega_j \omega_k \cos(\omega_j)\cos(\omega_k) \neq 0) \\
&= 0 \qquad \left(\text{since } \tfrac{\tan(\omega_j)}{\omega_j} = -(D/d - 1)\ (36)\right)
\end{aligned}
$$

single pattern case are shown in Figure 5.

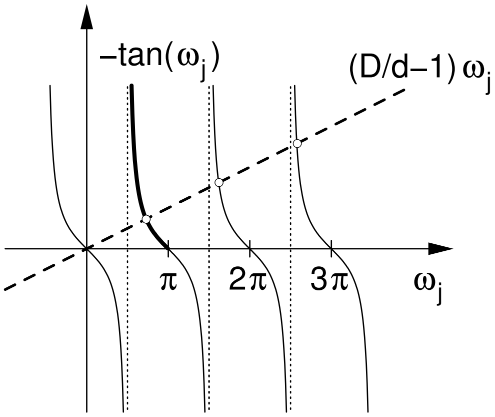

Figure 4: Illustration of the equation $\tan(\omega_j)/\omega_j = -(D/d-1)$ (36), which implicitly defines ω_j. Shown are four branches of $-\tan(\omega_j)$ and the line $(D/d-1)\omega_j$. The first intersection corresponding to ω_1 can lie on the section drawn with a thick line. Thus, it is obvious that $\pi/2 < \omega_1 \leq \pi$. Frequencies for higher odd indices are given by the other intersections with $\omega_j \in (j\pi/2, (j+1)\pi/2]$.

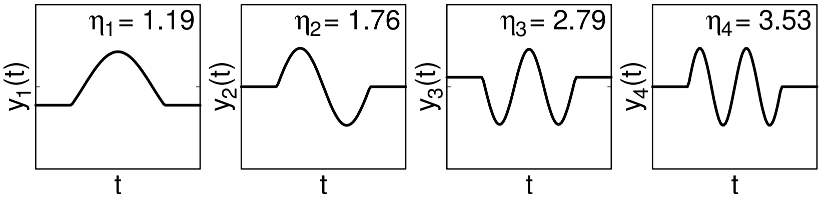

Figure 5: The first four exact Δ-optimal responses for the single pattern boundary condition. $D = 150$ and $d = 85$; t-axes range from t_A to t_B; y-axes range from -4 to $+4$. Notice, that in contrast to Figure 3 η_1 now has a realistic value greater than 1.

6.4 Multiple Patterns

Assume now that not only one but P different patterns are presented to the network at different non-overlapping time intervals. The responses must then be functions that may vary within the different time intervals in which a pattern is visible and are constant otherwise. What is the Δ-optimal set under these boundary conditions?

If we accept that the function set (33) is sufficiently exact, the answer is relatively simple. Since the time intervals of the different patterns are non-overlapping and the constant response is zero, it is obvious that the functions of the Δ-optimal sets of the different patterns, or different time intervals, are all mutually orthogonal and that also their time derivatives are. Thus, according to Corollary 2, a Δ-optimal set for the multiple pattern case is simply the union of the Δ-optimal sets of all the single patterns, with the functions ordered by their η-values; see Figure 6.

If the time intervals of the patterns differ, the η-values differ correspondingly according to (33), and the Δ-optimal set of functions is in general unique. If the time intervals are of similar but not identical length, the first unit responds with half a sine wave to the pattern visible within the longest interval, the second unit to the pattern visible within the second longest interval, etc. Then come units with full sine wave responses, first to the pattern visible within the longest interval, then to the pattern visible within the second longest interval, etc. The picture would be similar to Figure 6 but with slightly different intervals and η-values. If one pattern is presented for a much longer time interval than the others, leading to overall smaller η-values, the first few units may respond to this pattern, the first unit with half a sine wave, the second unit with a full sine wave, etc. Notice also that the response amplitude would be smaller for patterns presented for longer times.

In simulations like those shown in Figure 10 but with patterns of different size (15–30 units) and therefore different presentation times (70–85 time steps), neither of these two effects was significant. There was no clear tendency to represent large patterns only with early components and there was no negative correlation between pattern size and response strength. Instead the representation of each pattern was distributed over several output signal components as discussed in the next section. This indicates that the computational power of the network was not sufficient to reproduce this theoretically predicted behavior.

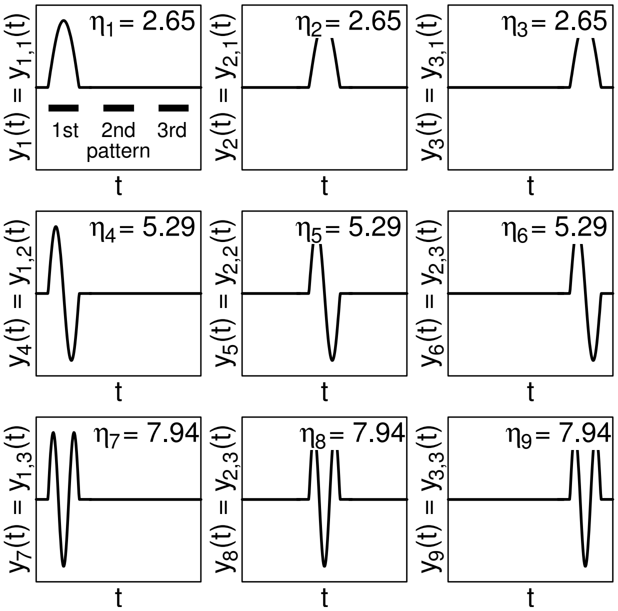

Figure 6: Approximate Δ-optimal set of responses for three patterns. Each column p forms a Δ-optimal set $\{y_{p,k}\}$ for a single pattern. United they form a Δ-optimal set $\{y_j\}$ for the corresponding three pattern case. However, this set is not unique, since the η-values in each row are identical; cf. Fig. 7. $D = 150$ and $d = 85$; t-axes cover an interval of length $3D$; y-axes range from -4 to $+4$.

6.4.1 Patterns of Same Duration

Consider now the case where the time intervals have identical length, so that for P time intervals there are P responses y_j with identical η-value, each one coming from a different pattern and therefore varying only in one interval and not the others, like in Figure 6. Focus on the responses with lowest η-value, which have the shape of half a sine wave, like y_1 in (33). Let $y_j(t_p)$ be the value of response y_j at some reference point of time interval p, in other words the response of unit j to pattern p at a certain location or size with $j, p \in \{1, ..., P\}$, and assume the responses are ordered such that unit j responds if $p = j$. The reference points are the same for all intervals and thus the response is the same for all patterns (in their respective interval) and denoted by r. The responses $y_j(t_p)$ then form a $P \times P$ diagonal matrix $\mathbf{Y}$ with all diagonal elements equal r. The responses of the P units to a single pattern p at the reference location form a *response vector* $(y_1(t_p), ..., y_P(t_p))^T$, which is the p-th column vector of matrix $\mathbf{Y}$. Since $\mathbf{Y}$ is a diagonal matrix, the response vectors for all P patterns are mutually orthogonal, which is convenient if one wants to recognize the patterns based on the responses.

We have seen above that since the η-values of the responses $y_j, j = 1, ..., P$ are identical, they are not unique. Any orthogonal transformation, written as an orthogonal $P \times P$ matrix $\mathbf{U}$, on the vector of these responses would yield an equally valid Δ-optimal set $y'_j(t) = \sum_{p=1}^{P} U_{jp} y_p(t)$ for $j = \{1, ..., P\}$; see Figure 7. The matrix of response vectors would change correspondingly and yield $\mathbf{Y}' = \mathbf{UY}$. In general each unit would then respond to each pattern to some extent, but still with half sine waves. However, the column vectors of $\mathbf{Y}'$ are still orthogonal. Thus the response vectors would still permit recognition of patterns equally well.

From another point of view one can also argue that the decorrelation constraint causes the row vectors of $\mathbf{Y}'$ to be orthogonal and the unit-variance constraint causes them to have identical norm, so that also the column vectors are orthogonal and the patterns can be recognized well. Thus the decorrelation constraint causes the system to generate different representations for different patterns.

In the simulation experiments (Wiskott & Sejnowski, 2002), if trained on a few patterns only, the response vectors were in fact close to orthogonal. For more training patterns, the network did not produce as many half sine wave responses as there were training patterns. Thus the response vectors could not all be mutually orthogonal. I found that in this situation the angle between any pair of response vectors would rarely be greater than 90°, indicating that the response vectors mostly lie within a cone of 90° opening angle or possibly within a rotated hyperquadrant. This also holds for testing patterns.

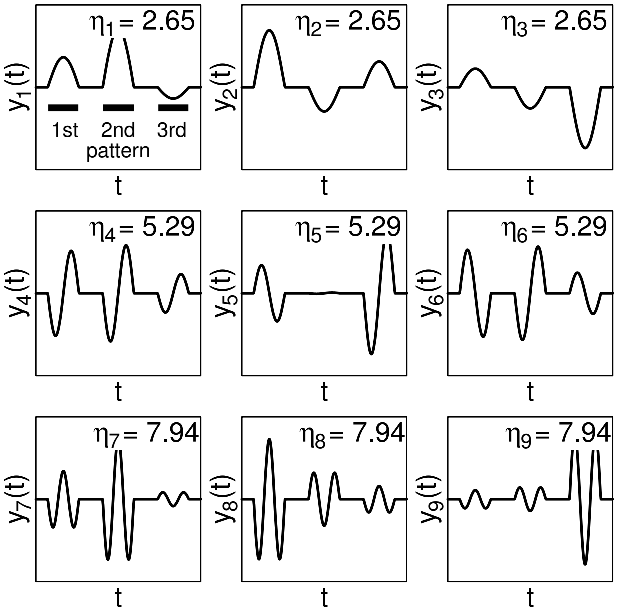

Figure 7: Approximate Δ-optimal set of responses for three patterns. Functions with identical η-value can be mixed by any orthogonal transformation and still form a Δ-optimal set. Notice that each of the three subsets belonging to a different η-value has been mixed with a different orthogonal transformation. $D = 150$ and $d = 85$; t-axes cover an interval of length $3D$; y-axes range from -4 to $+4$.

6.4.2 ***Where*-Responses**

The analysis given in the previous section is based on the assumption that the constant response during the intermissions is zero. This corresponds to the approximation made in Section 6.3 and holds for $D/d \rightarrow \infty$. The basic result is that in general all of the first P units should respond to all patterns with half a sine wave, but with different signs and amplitudes, so that the responses are uncorrelated and the response vectors are orthogonal. However, it is worth considering one particular response that appears often in simulations in more detail, namely the one that is (almost) identical for all patterns; see y_1 of Fig. 10. Such a response does not differentiate between different patterns and provides information only about their location or size. Thus it can be referred to as a *where*-response. The others, which differentiate between different patterns, are correspondingly referred to as *what*-responses. Notice that a *what*-response may still convey some *where*-information.

Since a half sine wave *where*-response is identical for all patterns, the optimal response is the one of the single pattern case simply duplicated P times and the corresponding η-value is multiplied by P. Thus we can use the exact function given by (35) (which is in fact not exactly half a sine wave). As one can infer in the limit $D \to d$ and can verify numerically, the η-value of such a response is in general higher than those with a zero constant response.

If we accept that such a half sine wave *where*-response is always generated, because it has a different η-value than the half sine wave *what*-responses, it is easy to see that the decorrelation constraint guarantees that the latter have zero mean exactly[9], although we have used the approximative form (33). Thus the constraints are fulfilled not only approximately but exactly, which may be taken as an *a posteriori* justification of the assumption of zero constant response. The first row in Figure 8 shows an example of these responses.

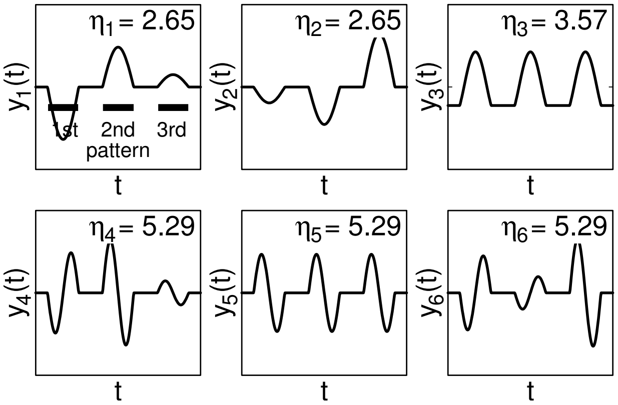

Figure 8: Δ-optimal set of responses for three patterns with two *where*-responses, y_3 and y_5. Notice that y_3 differs from y_1 and y_2 in its η-value and should therefore always occur as the third response. A response like y_5 occurs theoretically only by accident and could be any of the responses y_4, y_5, or y_6. $D = 150$ and $d = 85$; t-axes cover an interval of length $3D$; y-axes range from -4 to $+4$.

A *where*-response can also occur among the full sine wave responses. However, since it has the same η-value as the others, it should theoretically occur only by accident; see y_5 in second row in Figure 8. Taking together the first two *where*-responses provides a unique representation of location (remember that in case of size there are no full sine wave responses in the simulations for symmetry reasons, but the half sine wave response is sufficient to uniquely represent the size of a pattern). Even if these *where*-responses should not emerge explicitly, the corresponding information is usually still there and can be extracted by an appropriate orthogonal transformation.

6.4.3 Comparison with Simulation Results

To compare the theoretical results with simulation results we need to take into account that the simulations have the additional constraint that the responses are computed with some nonlinear functions from a given input signal. This causes the responses to be more irregular than predicted by the theory, which causes a shift of the η-values upwards. The amount of shift varies, however.

In the simulation for learning size invariance (Wiskott & Sejnowski, 2002, Fig. 21) the predicted η-value of the first *what*-responses is 10.08 and of the first *where*-response 12.99 (for 10 patterns with $D = 119$ and $d = 59$). The η-values of the first four simulation responses are in the range 10.3–11.31, indicating that the shift of η-values is small. Thus it is reasonable that the first responses are all *what*-responses, as predicted by theory.

In the simulation for learning translation invariance (Fig. 10) the predicted η-value for the first *what*-responses is 2.65, for the first *where*-response 3.57, and for all full sine wave responses (of *where*- as well as *what*-type) 5.29 (for 3 patterns with $D = 150$ and $d = 85$). Comparison with the first four simulation responses (cf. Fig. 10) indicates that the *what*-responses y_2 and y_4 suffer a significantly greater shift

[9] The inner product between the first *where*-response (y_3 in Fig. 8) and any of the first *what*-responses (y_1 and y_2 in Fig. 8) has no contribution from the intermissions, because there the *what*-responses are zero. Since during the individual pattern presentations the *where*-responses are identical and the *what*-responses only differ in amplitude but not in shape, the respective overall contribution can be written as a sum over the amplitudes of the *what*-responses times an integral over the *where*-response and a standard *what*-response. Due to the decorrelation constraint and since the integral is not zero, the sum over the amplitudes and therefore also the mean over the *what*-responses must vanish. Thus the zero-mean constraint is fulfilled exactly.

than the *where*-responses y_1 and y_3, presumably because the latter are easier to generate smoothly under pattern translation. As a consequence the half sine wave *where*-response comes first and the full sine wave *where*-response, like y_5 in Figure 8, does occur systematically and not only by accident, because it is now distinguished by its low η-value. However, it is not clear why it did not mix with half sine wave *what*-responses, like y_1 and y_2 in Figure 8.

With this, we can also understand why "for some parameter regimes, such as fewer patterns or smaller distances between patterns, no explicit *where*-components emerge" (Wiskott & Sejnowski, 2002, p. 743). Fewer patterns permit the system to generate smoother *what*-responses with correspondingly smaller η-values. Smaller distances between patterns cause shorter intermissions and a relative increase of the η-value of the half sine wave *where*-response. In both cases the half sine wave *where*-response is no longer distinguished by its low η-value and can mix with the half sine wave *what*-responses.

Figure 9, top row, shows four theoretically predicted responses similar in shape to and in the same order as those found in (Wiskott & Sejnowski, 2002, Fig. 11). Below are shown trajectory plots to provide a picture of the response vectors in phase space. Figure 10 shows corresponding simulation results from (Wiskott & Sejnowski, 2002, Fig. 11) for comparison.

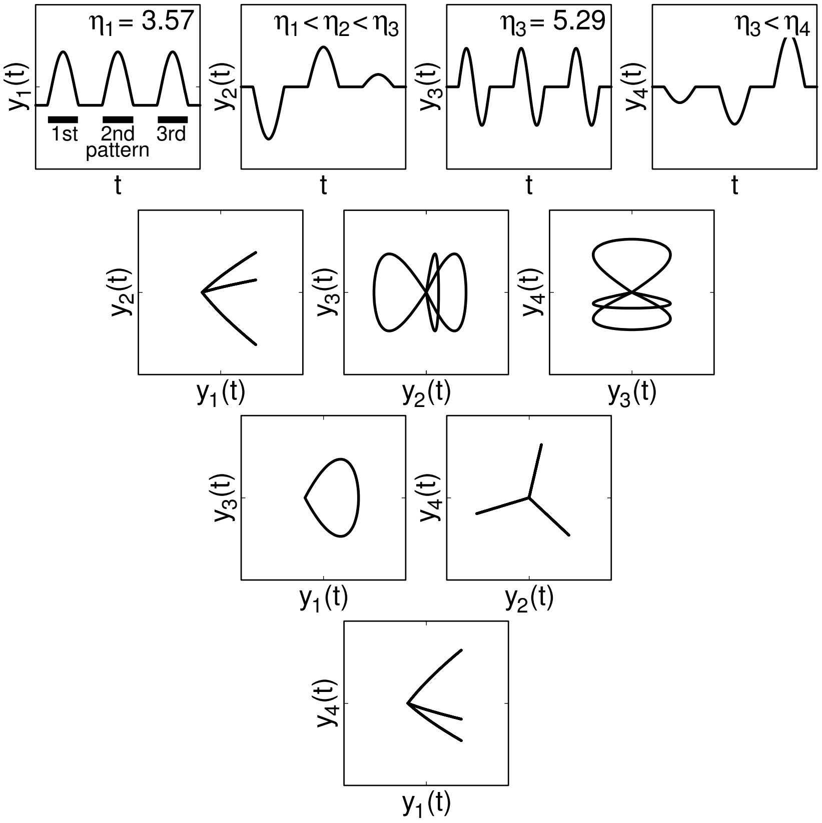

Figure 9: Four theoretically predicted responses arranged like the simulation results shown in Figure 10. η-values are estimated. $D = 150$ and $d = 85$ match the values in Figure 10. t-axes cover an interval of length $3D$; y-axes range from -4 to $+4$.

The qualitative agreement between the theoretically predicted responses and those obtained in the simulations is excellent. The simulation results are, of course, noisier and there are differences due to the arbitrary signs and amplitudes of responses to individual patterns. Another important difference is that the number of half sine wave responses the simulated network generated was much less than the number of patterns (= 20) presented during training, which is a consequence of the limited computational capacity of the network.

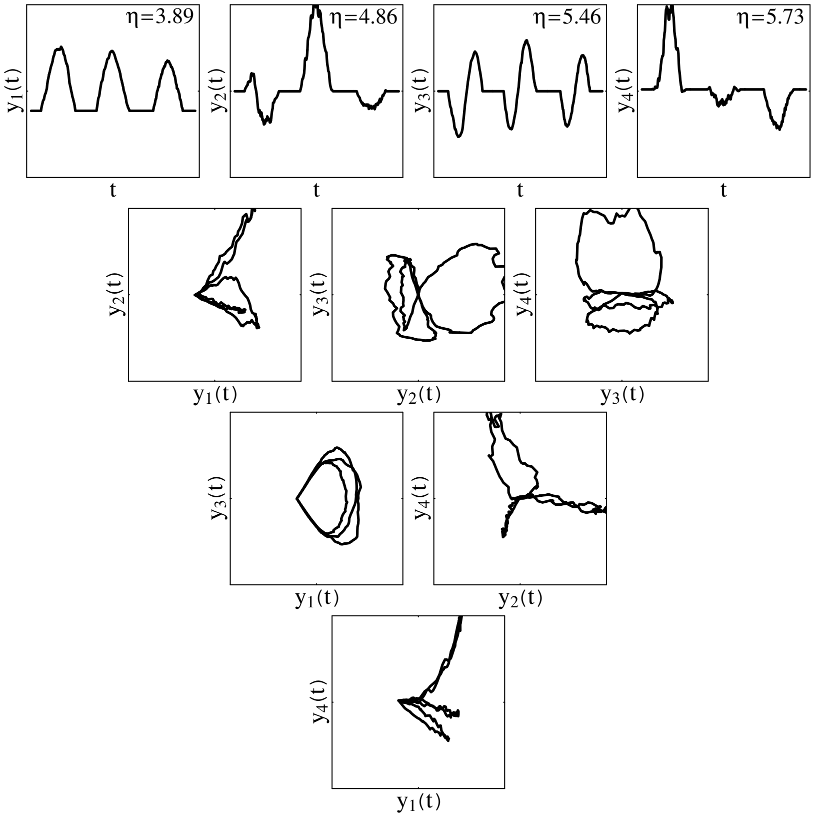

Figure 10: Simulation results from (Wiskott & Sejnowski, 2002, Example 4, Fig. 11). Shown are the results on the training data. $D = 150$ and $d = 85$. (In (Wiskott & Sejnowski, 2002, Fig. 11) patterns were actually presented for 84 time steps. After linear interpolation between these sample points the response may actually deviate from the constant response within a time interval of length 85. Thus, $d = 85$ and not 84.) t-axes cover an interval of length $3D$; y-axes range from -4 to $+4$. The graphs show only the response to three out of 20 patterns, thus the visible part of the responses do not fulfill the constraints (zero mean, unit variance, and decorrelation) exactly. Before computing the η-values the responses were normalized exactly.

7 Discussion

Slow feature analysis is in general a difficult variational calculus problem. We have seen that it can be idealized and simplified significantly if one abstracts from the input-output function and only considers the free output signal under some boundary conditions. This led to Optimization Problem 2. Solving this simplified optimization problem provides information about what the best solution is the system can obtain at all, regardless of the detailed constraints given by the input and the class of input-output functions. We were able to confirm several simulation results theoretically. General findings are:

- The responses tend to be sections of sine waves; see Eqs. (13, 29, 31, 33) and Figs. 9, 10.
- η-values usually increase linearly with the response index; see Eqs. (30, 32, 34). This also holds for the multiple pattern case, where the η-values of groups of responses increase linearly; see Fig. 6. The η-values are even comparable in magnitude to those of the single pattern case. For instance, the η-values in the third column of Figure 6 all coincide with the values of η_3, η_6, and η_9 for the single pattern case; see Eq. (34). Only the η-values of the first two columns are greater. This generally linear increase of the η-values has also been observed in simulation experiments ; see (Wiskott & Sejnowski, 2002, Figs. 5, 6, 7). However, the slope was much greater in these cases, because of the additional constraints given by the input signals and the class of input-output functions.

For the multiple pattern cases simulated in (Wiskott & Sejnowski, 2002, Examples 4, 5) and investigated here in Section 6.4 we find:

- The most invariant responses are not piecewise constant responses but half (or full) sine waves, since these have a less abrupt on- and off-set; see Figs. 9, 10 and (Wiskott & Sejnowski, 2002, p. 742).
- While individual *what*-responses vary like half sine waves and are therefore not truly invariant to translation or scale, the direction of the response vector, which is the vector of several responses at a given time, tends to be invariant over the pattern presentation; see Figs. 9, 10 (trajectory plots y_2 vs. y_1, y_4 vs. y_1, and y_4 vs. y_2).
- *Where*- as well as *what*-information is always extracted. If the η-values are sufficiently different *where*-responses emerge explicitly; see Figs. 9, 10 (y_1 and y_3). There are potentially many more *what*-responses than *where*-responses.
- Response vectors of different patterns tend to be orthogonal; see Sec. 6.4.1. This was also found in simulation experiments. If there were more patterns than *what*-responses, the response vectors were rarely more than 90° apart, indicating that they stay within a cone of 90° or a rotated hyperquadrant.
- If the duration of presentation is the same for all patterns, the representation of a particular pattern is in general distributed over all *what*-responses; see Figs. 9, 10 (y_2 and y_4) and (Wiskott & Sejnowski, 2002, Fig. 21) (all y_j). There is no tendency to generate a sparse representation. A representation in which only one unit responds at a time is even suppressed, because of the likely emergence of the first *what*-response, which enforces other units to respond with different signs to at least two patterns; see Sec. 6.4.1. However, this suppression is not strong and in the approximate analysis a representation in which only one unit responds at a time is possible; see Fig. 6 (y_1, y_2 and y_3). Thus, there may be room for an additional objective favoring sparse representations.

Furthermore, the theory would predict:

- If patterns are presented for different amounts of time, those seen longer are represented first; see Sec. 6.4. This would lead naturally to a sparse representation. However, this tendency was not observed in simulation experiments, indicating that it may be a weak effect or difficult to achieve in the network model considered here.

It is somewhat suspicious that the simulation results of Figure 10 could be reproduced so well without considering the input signal and the input-output function. However, this does not mean that the input is ignored or irrelevant. It rather indicates that the system has the tendency to generate the theoretically predicted responses if the input signal and the potential input-output functions allow it to do so. The response can always only be produced based on the input signal. But the theory leaves some room for adaptation to the input, because the responses of same type, half or full sine waves, can be mixed by any orthogonal transformation. A perfect fit of the simulation results with the theoretical responses, however, would clearly indicate overfitting. This is the case if only few training patterns are used, such as three or four. With 20 training patterns, like in the simulation results considered here, the effect of overfitting is weak and the system generalizes fairly well (cf. Wiskott & Sejnowski, 2002, Fig. 23). But then the fit is not perfect anymore. As indicated in Section 6.4.3, the simulation responses become noisier, have an unexpected order, and there are fewer than the expected 20 responses of the half sine wave type. Thus the simulations only reproduce part of the theoretically predicted output signal, but what they reproduce can be well understood theoretically.

Another issue we have touched upon only briefly is the question to what extent the learned representation is useful for recognition. There are two different aspects to discuss. One is the particular shape of individual response components, half and full sine waves, and the other is the population aspect of the response, such as questions of sparseness and orthogonality. As argued already in (Wiskott & Sejnowski, 2002) the half sine wave responses have the advantage over the more obvious piecewise constant responses that they avoid the abrupt on- and off-sets as a pattern moves into or out of the visual field. The full sine wave responses convey important *where*-information but do not seem to provide additional *what*-information. Thus in the simulations of (Wiskott & Sejnowski, 2002) only the full sine wave *where*-response was included in the analysis of results. On the population level, the tendency to produce orthogonal response vectors for different patterns is very useful for invariant pattern recognition. The overall orientation of the response vectors in space and with it the sparseness of the representation is theoretically very dependent on the exact training procedure. However, this effect was not observed in the simulations considered here. Experiments with more realistic input sequences are required for a more thorough evaluation.

In summary, the analysis presented here shows a way how simulation results obtained based on the principle of temporal slowness can be analyzed and understood theoretically. We have gained some general insights and a good understanding of responses obtained in (Wiskott & Sejnowski, 2002, Examples 4, 5). The analysis also shows that the learned representation depends strongly on the particular training procedure. For instance, it potentially makes a great difference whether patterns are presented for equal or different amounts of time.

Acknowledgments

I would like to thank Terrence Sejnowski and Andreas Herz for providing excellent working environments and general support, Michael Eisele for valuable suggestions in the beginning of this project, and Pietro Berkes and Raphael Ritz for critically reading the manuscript. At the Salk Institute for Biological Studies I have been supported by a Feodor-Lynen Fellowship from the Alexander von Humboldt-Foundation. At the Institute for Theoretical Biology I am supported by the Volkswagen-Foundation with a fellowship for a junior research group.

References

Becker, S. and Hinton, G. E. (1992). A self-organizing neural network that discovers surfaces in random-dot stereograms. *Nature*, 355(6356):161–163. 2

Berkes, P. and Wiskott, L. (2002). Applying slow feature analysis to image sequences yields a rich repertoire of complex cell properties. In Dorronsoro, J. R., editor, *Proc. Intl. Conf. on Artificial Neural Networks - ICANN'02*, Lecture Notes in Computer Science, pages 81–86. Springer. 2

Földiák, P. (1991). Learning invariance from transformation sequences. *Neural Computation*, 3(2):194–200. 2

Hinton, G. E. (1989). Connectionist learning procedures. *Artificial Intelligence*, 40(1-3):185–234. 2

Kayser, C., Einhäuser, W., Dümmer, O., König, P., and Körding, K. (2001). Extracting slow subspaces from natural videos leads to complex cells. In Dorffner, G., Bischoff, H., and Hornik, K., editors, *Proc. Intl. Conf. on Artificial Neural Networks - ICANN'01*, pages 1075–1080, Berlin, Heidelberg. Springer. 2

Mitchison, G. (1991). Removing time variation with the anti-hebbian differential synapse. *Neural Computation*, 3(3):312–320. 2

O'Reilly, R. C. and Johnson, M. H. (1994). Object recognition and sensitive periods: A computation analysis of visual imprinting. *Neural Computation*, 6(3):357–389. 2

Peng, H. C., Sha, L. F., Gan, Q., and Wei, Y. (1998). Energy function for learning invariance in multilayer perceptron. *Electronics Letters*, 34(3):292–294. 2

Stone, J. V. and Bray, A. J. (1995). A learning rule for extracting spatio-temporal invariances. *Network: Computation in Neural Systems*, 6(3):429–436. 2

Wallis, G. and Rolls, E. T. (1997). Invariant face and object recognition in the visual system. *Progress in Neurobiology*, 51(2):167–194. 2

Wiskott, L. and Sejnowski, T. J. (2002). Slow feature analysis: Unsupervised learning of invariances. *Neural Computation*, 14(4):715–770. 2, 6, 7, 9, 16, 17, 18, 19, 20, 21